Universidade de Aveiro Departamento de Matemática
2017

António da Costa Fernandes

Análise de conglomerados: comparação de técnicas e uma aplicação a dados de fluxo migratório em Portugal

Universidade de Aveiro Departamento de Matemática

2017

António da Costa Fernandes

Análise de conglomerados: comparação de técnicas e uma aplicação a dados de fluxo migratório em Portugal

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Doutora Adelaide de Fátima Baptista Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Dedico este trabalho à minha amada “ Aiza Alves” ao meu filho “Zanio”, aos meus pais e os irmãos, são a luz da minha vida e sem eles, nada seria possível.

O júri

Presidente

Prof. Doutor Pedro Filipe Pessoa Macedo
Professor Auxiliar do Departamento de Matemática, Universidade de Aveiro

Prof. Doutora Maria Fernanda Nunes Diamantino
Professora Auxiliar do Departamento Estatística e Investigação Operacional, Universidade de Lisboa

Prof. Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Auxiliar do Departamento de Matemática, Universidade de Aveiro (Orientadora)

Agradecimentos

Em primeiro lugar, gostaria de agradecer a Deus todo-poderoso pela saúde, força e disposição que sempre me proporcionou durante a elaboração desta dissertação, assim como, ao longo deste percurso académico.

À Profª Doutora Adelaide de Fátima Baptista Valente Freitas, como minha orientadora científica, pela sua paciência, dedicação e disponibilidade na orientação, e em ajudar-me a tirar todas as dúvidas ao longo do trabalho.

A todos os professores de Matemática em geral, e especificamente para os do mestrado em Matemática e Aplicações, por partilharem os seus conhecimentos científicos, pelos apoios, envolvimentos e disponibilidades a todas as unidades curriculares ao longo do estudo.

Ao Profº Doutor Domingos Cardoso, à Profª Doutora Clara Maria Magalhães, ao Dr. Ângelo Ferreira e ao Eng. Miguel de Oliveira pelos seus máximos apoios.

À Profª Doutora Maria Cristina Gomes, do Departamento de Ciências Sociais, Políticas e de Território da Universidade de Aveiro pelo apoio na disponibilização dos dados.

Aos meus amigos compatriotas Timorenses que estão nesta Universidade, pelos seus apoios e motivações. A todos os amigos do curso pelas suas amizades e compreensões durante a minha ausência ao longo deste percurso académico. À minha família, pelo grande apoio, compreensão, carinho, amor e pelo estímulo constante à conclusão deste trabalho.

À Universidade de Aveiro, em particular, ao Departamento de Matemática pela oportunidade e pelas excelentes condições proporcionadas para desenvolver este trabalho. À Universidade Nacional Timor Lorosa'e na parceria com Universidade de Aveiro, no sentido da cooperação continuar para capacitar os docentes.

Palavras-chave Conglomerado, Dissimilaridade e Similaridade, Técnicas aglomerativas, Técnicas divisivas, Fluxo migratório.

Resumo

A análise de conglomerados é um procedimento de Estatística Multivariada que tem como objetivo identificar e, eventualmente em consequência, classificar objetos ou traços de indivíduos de modo a que, dentro de um mesmo grupo os elementos sejam o mais homogéneo possível e entre grupos o mais heterogéneo.

O objetivo geral deste trabalho consiste na comparação de técnicas de conglomerados. Em particular focam-se: i) a construção de agrupamentos na Análise de Conglomerados usando técnicas hierárquicas de métodos aglomerativos, ii) o uso do coeficiente de correlação cofenética na comparação de resultados da aplicação de cada técnica estudada no processo de aglomeração (agrupamento/*clustering*), iii) os métodos de validação interna e de estabilidade, que têm como objetivo avaliar as qualidades e estabilidades dos grupos (*clusters*) obtidos por distintos métodos de aglomeração e, por fim, iv) uma aplicação dos métodos e procedimentos estudados a uma base de dados reais relativa ao fluxo migratório entre distritos em Portugal entre o ano 2005 e 2011, em termos de sexo, idade, situação de trabalho e habilitação académica; comparam-se os resultados da aplicação de cada técnica estudada no processo de aglomeração usando a correlação cofenética e avaliam-se os grupos obtidos usando medidas de validação interna e de estabilidade.

Relativamente ao conjunto de dados estudado verificou-se, pelos resultados das validações, que a maioria dos métodos hierárquicos com o critério de aglomeração da média apresentam melhores indicadores em termos de correlação cofenética, validação interna e estabilidade dos grupos formados quando analisadas as diferentes características separadamente. Agregando toda a informação numa só matriz de dados, verificou-se que tal método hierárquico exibe qualidade, mas não estabilidade dos grupos.

Keywords Cluster, Dissimilarity and similarity, Agglomerative method, Divisive method, Migration flows.

Abstract Cluster analysis is a method of multivariate statistical analysis aimed at identifying and, consequently, classifying objects or individuals such that elements in the same cluster are as homogeneous as possible and in different clusters are the most heterogeneous.

The main goal of this dissertation is the comparison of clustering techniques. In particular, it is intended to focus on: i) construction of clusters using agglomerative hierarchical techniques, ii) use of cophenetic correlation for comparison of results provided by different linkage criteria, iii) use of internal validation and stability methods in order to evaluate the qualities and stability of the clusters obtained by different methods of Cluster Analysis, and finally, iv) an application of the studied methods to a real data set on migration flow in Portugal between 2005 and 2011; several features like gender, age, work situation and academic qualification will be analyzed. The evaluation of the clusterings will be performed using cophenetic correlation, internal validation and stability measures.

Considering the data set, the hierarchical method with average linkage criterion leads to the best clustering in terms of internal validation and stability measures when different characteristics are analyzed separately. Aggregating all the information into a single data matrix, the clustering provided by the linkage criterion exhibits quality (internal validation) but not stability.

Conteúdos

Conteúdos i
Lista das figuras iii
Lista das tabelas iv
Abreviaturas v
Capítulo 1 Introdução
1.1. Contextualização 1
1.2. Objetivos 2
1.3. Metodologia da investigação 3
1.4. Estrutura da investigação 3
Capítulo 2 Enquadramento teórico
2.1. Metodologia de Análise de Clusters 5
2.1.1. Seleção das variáveis 5
2.1.2. Medidas de proximidade 6
2.2. Métodos hierárquicos 16
2.2.1. Definição 16
2.2.2. Métodos aglomerativos 17
2.2.2.1. Método do vizinho mais próximo (single linkage) 18
2.2.2.2. Método do vizinho mais distante (complete linkage) 21
2.2.2.3. Método de média (average linkage) 22
2.2.2.4. Método centróide 24
2.2.2.5. Método de Ward 27
2.2.3. Métodos divisivos 33
2.2.3.1. Métodos divisivos monotéticos 34
2.2.3.2. Métodos divisivos politéticos 37
2.3. Comparação de métodos aglomerativos 41
2.4. Métodos não-hierárquicos 45
2.5 Medidas de validação de clusters 54
2.5.1. Medida de validação interna 54
2.5.2. Medida de estabilidade dos Clusters 60
Capítulo 3 Aplicações e análise de resultados
3.1. Descrição dos dados 69
3.2. Análise e comparação de técnicas hierárquicas 70

3.3. Métodos de partição 78
3.4. Validação interna e validação de estabilidade 82
Capítulo 4 Considerações finais
4.1. Conclusão 95
4.2. Sugestões 96
Bibliografia 97
Anexo A: Conjunto de dados reais analisados 99
Anexo B : Scripts do R 99

Lista das figuras

Figura 1: Gráfico representando área versus número de chefes de família 9
Figura 2:Dendrograma do critério do vizinho mais próximo do Exemplo 4 20
Figura 3:Dendrograma usando o método do vizinho mais afastado do Exemplo 5 22
Figura 4:Dendrograma usando o método Average do Exemplo 6 24
Figura 5:Dendrograma usando o método de centróide do Exemplo 7 26
Figura 6: Dendrograma obtido pelo método de Ward do Exemplo 9. 32
Figura 7: Gráfico ilustrando a relação de procedimento entre um método aglomerativo e um método divisivo. 34
Figura 8:Exemplo do método divisivo Diana para o Exemplo 11 41
Figura 9: Dendrograma pelo método da média sobre a característica Idade 72
Figura 10: Dendrograma pelo método da média sobre a característica Género 73
Figura 11: Dendrograma pelo método da média sobre a característica Situação do trabalho 74
Figura 12: Dendrograma pelo método da média sobre a característica Habilitação literária 75
Figura 13:Dendrograma pelo método da média sobre as quatro características em conjunto. 76
Figura 14: Gráficos para investigar o número de clusters a considerar no método k-médias 78
Figura 15:Gráfico de dispersão para a característica género, diferenciando os clusters encontrados pelo método k-médias com k=2 79
Figura 16:Gráfico da aplicação do método silhueta com k=2 para os dados gerais 81
Figura 17:Gráfico validação interna dados idade 83
Figura 18:Gráfico de validação da estabilidade para os dados idade 84
Figura 19:Validação interna para os dados situação do trabalho 86
Figura 20:Validação da estabilidade para os dados situação do trabalho 87
Figura 21:Validação interna dos dados habilitação literária 89
Figura 22:Validação estabilidade para dados habilitação literária 90
Figura 23:Validação interna para os dados gerais 91
Figura 24:Validação da estabilidade para os dados gerais 92

Lista das tabelas

Tabela 1: Dados do Censo de 2015 de Timor-Leste 8
Tabela 2: Distâncias intercidades de Timor-Leste 12
Tabela 3: Tabela de contingência de resultados binários para dois indivíduos..... 14
Tabela 4: Coeficientes de semelhança mais usados para dados binários 14
Tabela 5: Matriz de dados 18
Tabela 6: Parâmetros de Lance-Williams para vários métodos hierárquicos..... 33
Tabela 7: Tabela contingência num par de variáveis binárias nos métodos monotéticos. .. 35
Tabela 8: Interpretação do coeficiente de silhueta (CS)..... 58
Tabela 9: Identificação das variáveis relativas aos dados estudados..... 70
Tabela 10: Resultado do coeficiente correlação cofenética dos métodos hierárquicos aplicados sobre os dados 71
Tabela 11: Tabela resumo dos clusters obtidos aplicando o método divisivo DIANA..... 77
Tabela 12:Resultados do método k-medóides através do método das silhuetas 80
Tabela 13:Validação interna dos dados idade assinalando o número de clusters k onde o valor ótimo da medida foi atingido e qual o método para qual tal resultou. 82
Tabela 14:Validação da estabilidade dos dados idade assinalando o número de clusters k onde o valor ótimo do índice foi atingido e qual o método para qual tal resultou 84
Tabela 15:Validação interna dos dados situação do trabalho assinalando o número de clusters k onde o valor ótimo do índice foi atingido e qual o método para qual tal resultou 85
Tabela 16:Validação da estabilidade para os dados situação do trabalho 86
Tabela 17:Validação interna para os dados habilitação literária 88
Tabela 18:Validação de estabilidade para dados habilitação literária..... 89
Tabela 19:Validação interna para os dados gerais..... 91
Tabela 20:Validação de estabilidade dados geral..... 92

Abreviaturas

$\boldsymbol{d_{ij}}$	Distância entre os objetos i e j
$\boldsymbol{s_{ij}}$	Medida similaridade entre os objetos i e j
$\boldsymbol{d_{AB}}$	Distância entre os *clusters* A e B
CD	Coeficiente de divisivos
PAM	*Partitioning Around Medoid* (Método de partição de medóides)
CS	Coeficiente de silhuetas
APN	*Average proportion of non-overlap* (Proporção média de não sobreposição)
AD	*Average distance* (Distância média)
ADM	*Average distance between means* (Distância média entre médias)
FOM	*Figure of merit* (Figura de mérito)

Capítulo 1

Introdução

Neste primeiro capítulo é apresentada uma breve contextualização do tema, o objetivo do trabalho, a metodologia da investigação e, por fim, a estrutura global da presente dissertação.

1.1. Contextualização

No dia-a-dia falamos de agrupamentos em inúmeros aspetos da nossa vida. Por exemplo, no conjunto dos alimentos, os vegetais, as frutas, os lacticínios, etc. Na verdade, as pessoas aprendem a classificar os objetos pertencentes ao seu ambiente envolvente, e a associar os resultados dessa classificação por palavras da sua linguagem. Em contextos mais gerais, a identificação de agrupamentos de indivíduos similares caracterizados por múltiplos traços é muito utilizada em muitas áreas científicas, desde as Engenharias, às Ciências Sociais, Ciências da Saúde ou Ciências Experimentais, entre outras.

A análise de conglomerados também habitualmente designada por Análise de *Clusters* é um procedimento de Estatística Multivariada que tem como objetivo identificar e, em consequência, classificar objetos ou traços de indivíduos de modo que, dentro de um mesmo grupo os elementos sejam o mais homogéneo possível e entre grupos o mais heterogéneo.

A análise multivariada corresponde à análise estatística mais usada para analisar dados definidos por várias variáveis podendo serem ou não mutuamente correlacionadas entre si.

Na Análise de *Clusters* de indivíduos, os possíveis agrupamentos realizados mantêm o carácter multivariado que descreve os indivíduos, o que não sucede em outras técnicas multivariadas que têm como objetivo a redução da dimensionalidade dos dados (exemplo: análise de componentes principais).

Vários autores dividem as técnicas de análise multivariada em duas classes, nomeadamente, as técnicas de dependência e as técnicas de interdependência. Na análise de

dependência, uma ou mais algumas variáveis dependem de outras variáveis, como, por exemplo, análise de regressão linear multivariada, análise discriminante, a análise de variância (MANOVA), e análise de correlação canónica. Na análise de interdependências todas as variáveis não dependem uma das outras como é, por exemplo, o caso da análise fatorial, Análise de *Clusters*, etc.

A análise de agrupamentos estuda todo um conjunto de relações interdependentes entre variáveis ou entre indivíduos. Ela não faz distinção entre variáveis dependentes e independentes, isto é, não existe uma relação do tipo causa e efeito como na regressão. Nesta dissertação, em termos de aplicação, dá-se particular enfoque ao agrupamento de indivíduos, investigando a existência de relações de interdependência ou grupos de distritos de Portugal com comportamentos distribucionais similares, em termos de distintas características, nomeadamente, sexo, idade, situação de trabalho e habilitação académica.

1.2. Objetivos

Existem diversas técnicas para identificar ou encontrar agrupamentos. Uma das técnicas mais usadas são os métodos hierárquicos e os métodos de partição. Para ambos é necessário definir uma medida de proximidade entre objetos ou traços de objetos. Para os métodos hierárquicos é necessário definir o critério de agregação de grupos. Nesta dissertação, o objetivo desta pesquisa é melhorar o conhecimento do pesquisador sobre técnicas de análise de conglomerados e compará-las. Em particular, daremos destaque aos métodos hierárquicos estudando diversos critérios de agregação considerados na literatura especializada.

Especificamente, os objetivos a serem alcançados nesta pesquisa, são:

1. Ilustrar como são construídos os agrupamentos na Análise de *Clusters* usando técnicas hierárquicas e outras;
2. Considerar o uso da correlação cofenética na comparação de resultados da aplicação de cada técnica estudada no processo de aglomeração (agrupamentos/*clusterings*);
3. Mostrar métodos de validação interna e de estabilidade que têm como objetivo avaliar as qualidades e estabilidades dos *clusters* obtidos pelos distintos métodos de Análise de *Clusters* ;

4. Aplicar os métodos estudados a dados reais de fluxo migratório no território de Portugal entre o ano 2005 e 2011 e comparar os resultados da aplicação de cada técnica estudada no processo de aglomeração (agrupamento).

1.3. Metodologia da investigação

Para este estudo começaremos por uma revisão da literatura especializada com vista a recolher informação sobre vantagens e desvantagens conhecidas de diferentes métodos de aglomeração e medidas de proximidade. O objetivo é contribuir para a construção de um estudo teórico sobre técnicas multivariadas de análise de conglomerados. A seguir iremos aplicar as metodologias estudadas a dados, quer contidos na base de dados do *software* R quer outras. Com recurso ainda ao R, iremos aplicar as metodologias estudadas aos dados e efetuar comparações de resultado do trabalho e obter conclusões.

1.4. Estrutura da investigação

Esta investigação está desenvolvida em quatros capítulos. No primeiro capítulo efetuou-se a introdução da investigação; é constituído pela introdução, objetivo, metodologia da investigação e a estrutura da investigação. O segundo capítulo foca-se no enquadramento teórico e é constituído por uma síntese dos trabalhos que se consideram mais relevantes para a definição do quadro teórico subjacente à investigação. O terceiro capítulo ilustra aplicações e análises a dados reais relativos aos fluxos migratórios entre distritos em Portugal entre o ano 2005 e 2011. E finalmente no quarto capítulo faz-se uma análise dos resultados encontrados e incluem-se as principais conclusões e implicação da investigação, bem como sugestões para trabalho futuro.

Capítulo 2

Enquadramento teórico

Neste capítulo é apresentado o enquadramento teórico de Análise *de Clusters*, como a metodologia de Análise de *Clusters*, descrevendo as seleções de variáveis e as medidas de proximidade. É também exposta uma breve descrição de métodos hierárquicos e não hierárquicos. E, por fim, apresentam-se as comparações de dendrogramas e as validações de *clusters* através de validação interna e validação de estabilidade de *clusters*.

2.1. Metodologia de Análise de *Clusters*

Segundo Reis (2001) uma Análise de *Clusters* de indivíduos processa-se de acordo com o seguinte procedimento:

- Seleção dos indivíduos ou de uma amostra de indivíduos a serem agrupados;
- Seleção das variáveis ou definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
- Definição de uma medida de similaridade ou de distância entre dois indivíduos;
- Escolha de um critério de agregação ou desagregação de grupos de indivíduos, isto é, a definição de um algoritmo de classificação ou de partição;
- Por último, a validação dos resultados.

2.1.1. Seleção das variáveis

Na Análise de *Clusters* de indivíduos é fundamental ter particular cuidado na seleção das variáveis de partida que vão caracterizar cada indivíduo e determinar, em última instância, qual o grupo em que deve ser inserido. Um dos itens importantes na Análise de *Clusters* é a seleção das variáveis que serão utilizadas para o agrupamento (formação de *clusters*).

Segundo Reis (2001), a seleção de variáveis para uma Análise de *Clusters* contém duas questões prévias associadas: por um lado, o investigador deve indicar as características fundamentais do problema concreto em estudo, e por outro lado, deverá ter-se em conta a

escala de medidas dessas características. Assim, a seleção das variáveis é um dos aspetos que pode influenciar os resultados da Análise de *Clusters*, pelo que o objetivo da aplicação de técnicas de agrupamento não pode ser separada de um estudo prévio sobre a seleção de variáveis a usar para classificar os objetos em *clusters*. A seleção de variáveis deve estar de acordo com as teorias e conceitos que são comummente usados e têm de ser racional.

As únicas variáveis selecionadas que podem caracterizar os objetos a serem agrupados, e especificamente, devem estar em conformidade com a finalidade da Análise de *Clusters*.

2.1.2. Medidas de proximidade

Medidas de proximidade são medidas quantitativas usadas para representar a proximidade de objetos ou de variáveis, consoante se pretende realizar agrupamento de objetos ou agrupamento de variáveis, respetivamente. Uma medida de proximidade representa uma medida de similaridade se o valor da medida aumenta à medida que dois elementos estão mais próximos ou são mais similares. Uma medida de proximidade representa uma medida de dissimilaridade ou de distância se o valor da medida diminui à medida que dois elementos se tornam mais parecidos ou menos distantes (Timm, 2002).

Em relação a escolher uma medida de proximidade, quer seja uma medida de dissimilaridade (distância) ou de similaridade a aplicar aos dados, não existe consenso sobre uma fórmula universal para tal; por isso, investigadores continuam a propor novos coeficientes.

Gower & Legendre (1986, pp.31) referiram "um coeficiente tem de ser considerado no contexto do estudo estatístico, incluindo a natureza dos dados e do tipo de análise pretendido". Alertam ainda que a matriz dos dados, a escala de medição usada para avaliar as características de interesse de cada indivíduo em observação (escala de medida nominal, ordinal, intervalar ou razão), a natureza do problema (aglomerar variáveis ou aglomerar unidades amostrais) e o método a aplicar na obtenção dos agrupamentos devem influenciar a escolha da medida de proximidade.

A medição é um processo que determina um número ou atributo associado a determinadas características de acordo com regras ou procedimentos pré-determinados. Por

exemplo, os indivíduos podem ser descritos com respeito a um número de características, tais como o grupo etário (criança, jovem, adulto ou idoso), género (masculino ou feminino), número de irmãos, altitude da sua residência atual. Todas as escalas de medição podem ser classificadas em quatro tipos, nomeadamente: escala nominal, escala ordinal, escala de intervalo e escala de razão (Hand, 1996; Stevens, 1946; Sharma, 1996).

A escala nominal pode ser usada para classificar variáveis. Uma variável nominal é uma medida de categorias[1].

A escala ordinal mantém as propriedades da escala nominal, mas os seus atributos qualitativos são ordenáveis pelo que tem a capacidade de ordenar os dados. Por exemplo, nível de satisfação de um cliente, grau académico, grupo etário.

A escala intervalar é uma medida que contém todas as características da escala ordinal e além de isso conhecem-se as distâncias entre quaisquer dois números (posição) desta escala[2]. O valor zero é um número como outro qualquer, por exemplo: Temperatura igual a 0º significa que há temperatura, sendo aquele estado atribuído o valor de 0.

A escala de razão é uma medida que tem todas as caraterísticas das escalas discutidas anteriormente, mas fornece um zero absoluto ou uma origem significativa (sendo que o zero significa ausência de atributo), permitido calcular o cociente entre dois valores. Exemplo: peso, altura, comprimento.

Consideremos um conjunto de n indivíduos[3] descritos por p características ou variáveis. Este conjunto pode ser representado sob a forma de uma matriz X, de dimensão $n \times p$, em que as n linhas correspondem aos indivíduos e p corresponde às colunas de variáveis, dada por:

[1] Exemplo escala nominal; género, cores, tipo de sangue, cidade, tipo de doença, religião.

[2] Exemplo: a temperatura de 50°c é mais elevada do que 40ºc, de igual modo que a temperatura de 30ºc é mais elevada do que 20ºc. A diferença de temperaturas 50ºc - 40ºc é igual à diferença 30ºc – 20ºc, e é de 10ºc. Assim, podemos distinguir (categorizar) o tipo de valor também podemos calcular a distância ou intervalo.

[3] que podem ser as pessoas, animais, plantas, empresas, países ou mesmo palavras, entre outas entidades.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

O elemento x_{ij} representa o atributo da variável j para o objeto (indivíduo) i.

Exemplo 1: Timor-Leste está dividido em treze municípios administrativos. Em períodos predefinidos, o governo de Timor-Leste determina a recolha dados através de censos. A tabela de dados seguinte apresenta dados recolhidos no Censo de 2015 sobre a área de cada município (em km^2) denotada por x_1, o número total de chefes das famílias (x_2) e o número total de pessoas em cada município (x_3).

No	Municípios	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$
1	Aileu	737	7832	48554
2	Ainaro	804	2819	66397
3	Baucau	1506	23195	124061
4	Bobonaro	1376	18192	98932
5	Covalima	1203	13285	64550
6	Díli	367	39310	252884
7	Ermera	768	21069	127283
8	Lautem	1813	11969	64135
9	Liquiçá	549	12800	73027
10	Manatuto	1782	7796	45541
11	Manufahi	1323	9257	52246
12	Oecusse	814	15131	72230
13	Viqueque	1877	15589	77402

Tabela 1: Dados do Censo de 2015 de Timor-Leste

A Figura 1 ilustra o diagrama de dispersão associado ao par de variáveis (x_2, x_1). Verificamos que, quando projetamos os pontos (cidades) sobre o eixo vertical, é o município de Díli que apresenta maior número de chefes da família e menor de área, comparado com outros municípios que possuem maiores áreas e menores números de chefes de famílias.

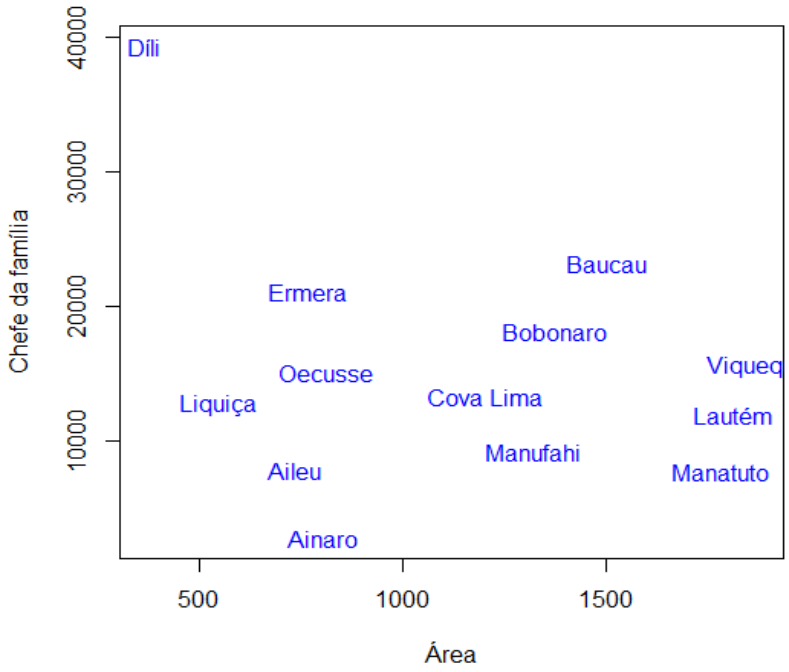

Figura 1: Gráfico representando área versus número de chefes de família

2.1.2.1. Medidas de dissimilaridade entre objetos

Os diagramas de dispersão permitem visualizar a maior ou menor distância entre objetos bidimensionais. Quando se pretende avaliar a dissimilaridade entre objetos p-dimensionais, $p > 3$, não é possível visualmente avaliar a proximidade entre objetos, pelo que será necessário estender a noção intuitiva de distância.

Medidas de dissimilaridade e distâncias para dados quantitativos

Dado dois objetos, num espaço p-dimensional, descritos pelas linhas i e j da matriz X, uma medida de dissimilaridade d_{ij} representa uma medida de distância entre os indivíduos i e j e satisfaz as seguintes condições:

- $d_{ij} \geq 0$
- $d_{ij} = 0$ se e só se $i = j$
- $d_{ij} \leq d_{ik} + d_{kj}$

- $d_{ij} = d_{ji}$

onde a primeira condição implica que a medida não é negativa, a segunda condição exige que a medida seja nula quando o objeto i é igual a j, a terceira condição é a desigualdade triangular e, finalmente, a quarta condição implica que a medida é simétrica.

Apresentam-se, em seguida, algumas medidas de distância, d_{ij}, de acordo com vários autores (Everitt, 2011; Reis, 2001;Rencher, 2012; Johnson *et al*, 2014) :

Distância Euclidiana é definida pela expressão:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Quadrado da distância Euclidiana é definida pela seguinte expressão:

$$d_{ij}{}^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$$

Distância absoluta ou *City-Block metric* é definida pela seguinte expressão:

$$d_{ij} = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$$

Distância de Minkowski é definida a partir da medida anterior; se $r \geq 1$, é a distância absoluta; se $r = 2$ é a distância Euclidiana. A expressão, em termos de r, é:

$$d_{ij} = \left(\sum_{k=1}^{p}|x_{ik} - x_{jk}|^r\right)^{1/r}$$

Distância de Mahalanobis é definida, ao contrário das anteriores, à custa da matriz de covariâncias $\sum$ e é dada por:

$$d_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \sum^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)$$

onde $\boldsymbol{x}_i = [x_{i1}\ x_{i2}\ \ldots\ x_{ip}]^T$ e $\boldsymbol{x}_j = [x_{j1}\ x_{j2}\ \ldots\ x_{jp}]^T$ representam os vetores com os valores das variáveis para os indivíduos i e j.

Distância de *Chebyshev* é definida pelo valor máximo, para todas as variáveis, das diferenças entre os dois indivíduos, i.e.,

$$d_{ij} = \max_k |x_{ik} - x_{jk}|$$

Exemplo 2: Relativamente aos dados do Exemplo 1, calculou-se a distância entre os distritos Aileu e Ainaro tomando diferentes medidas:

Medida de distância	Valor da distância
Distância Euclidiana	18533,95
Quadrado da distância Euclidiana	343507303
Distância absoluta ou City-Bloc metric	22,923
Distância de Mahalanobis	0,34
Distância de Chebyshev	17843

Escolhida uma medida de distância, a informação sobre a distância entre (n) objetos pode ser apresentada por uma matriz D de dimensão $n \times n$ cujos elementos medem a proximidade (distância) entre cada par de indivíduos, ou seja

$$D = \begin{bmatrix} 0 & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & 0 \end{bmatrix}$$

em que d_{kl} é a distância entre os objetos k e l, com k e $l = 1,2,...,n$.

Exemplo 3: Timor-Leste é composto por 13 municípios. A matriz de distâncias intercidades de municípios é dada na Tabela 2. Aqui, a medida de distância considerada corresponde à distância rodoviária entre cidades.

A matriz de distâncias pode ser considerada mais genericamente tomando uma medida de proximidade (de dissimilaridade ou de similaridade), e, consequentemente, D ser designada por matriz de proximidade. Para a sua construção é necessário previamente selecionar a medida de proximidade a usar.

No.	Município	Cidade	Distâncias intercidades (em km^2)												
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	Ainaro	Ainaro	0	66,2	209,8	89,7	67	110,3	93,9	138,8	275,8	56,7	138,4	223,9	151,3
2	Aileu	Aileu	66,2	0	166	100	141,4	44,2	27,8	72,2	311,4	69	108,3	230,9	186,9
3	Baucau	Baucau	209,8	166	0	258	235,3	125,5	167,4	155,7	89,2	176,4	58,8	328,3	58,5
4	Bobonaro	Maliana	89,7	100	258	0	74,8	116,2	73,4	120,2	324	104,8	181,9	136,2	199,5
5	Covalima	Suai	67	141,4	235,3	74,8	0	157,6	114,8	165	301,3	82,2	199,8	178,4	176,8
6	Díli	Díli	110,3	44,2	125,5	116,2	157,6	0	44	32,3	304,9	204,7	67,3	205	180,4
7	Ermera	Gleno	93,9	27,8	167,4	73,4	114,8	44	0	51,4	357,5	96,8	109,6	204,3	233
8	Liquiçá	Liquiça	138,8	72,2	155,7	120,2	165	32,3	51,4	0	335,6	141,6	98	172,8	211,1
9	Lautém	Lospalos	275,8	311,4	89,2	324	301,3	304,9	357,5	335,6	0	242,4	146,2	458,2	127,4
10	Manufahi	Same	56,7	69	176,4	104,8	82,2	204,7	96,8	141,6	242,4	0	141	239,1	117,9
11	Manatuto	Mantuto	138,4	108,3	58,8	181,9	199,8	67,3	109,6	98	146,2	141	0	270,6	116,7
12	Oecusse	Pante Macassar	223,9	230,9	328,3	136,2	178,4	205	204,3	172,8	458,2	239,1	270,6	0	334,8
13	Viqueque	Viqueque	151,3	186,9	58,5	199,5	176,8	180,4	233	211,1	127,4	117,9	116,7	334,8	0

Tabela 2: Distâncias intercidades de Timor-Leste

2.1.2.2. Medidas de similaridade entre objetos

A maioria dos métodos de Análise de *Clusters* requer uma medida de dissimilaridade entre os elementos a serem agrupados, normalmente expressa como uma distância ou métrica. Contudo, por vezes, em vez de se usar uma medida de dissimilaridade d_{ij} para avaliar a distância entre dois objetos i e j, é aplicada uma medida de similaridade s_{ij}. Nesse caso, quando os objetos i e j são iguais, o valor s_{ij} de similaridade torna-se máxima. Tipicamente, uma medida de similaridade varia entre 0 e 1, em que 0 significa que os dois objetos não são semelhantes e 1 reflete a similaridade máxima.

Uma medida de similaridade, s_{ij} entre os objetos i e j, caracteriza-se pelas seguintes propriedades (Tim, 2002; Kaufman *et al*, 2005):

- $0 \leq s_{ij} \leq 1$
- $s_{ii} = 1$
- $s_{ij} = s_{ji}$

onde a terceira condição implica que a medida é simétrica, enquanto que as duas primeiras garantem que a medida é sempre positiva e toma o valor máximo (1) se e somente se os objetos i e j forem idênticos.

Se os dados relativos a um conjunto de dados forem apresentados por uma matriz de similaridade, podemos transformar as similaridades em dissimilaridades. A dissimilaridade d_{ij} pode-se obter da similaridade s_{ij} usando uma função decrescente de s_{ij}, com $s_{ij} \geq 0$. Desse modo, quanto maior a similaridade s_{ij} entre i e j, menor a sua dissimilaridade d_{ij} deve ser. Por exemplo, podem considerar-se as transformações $d_{ij} = 1 - s_{ij}$ ou $d_{ij} = \sqrt{1 - s_{ij}}$. Estas transformações tornam as diferenças entre as grandes semelhanças mais importantes, mas por outro lado, torna mais difícil a obtenção de pequenas dissimilaridades (Kaufman at al, 2005). Como consequência, a matriz de dissimilaridade resultante pode ser bastante homogénea e menos suscetível de produzir agrupamentos nítidos.

Medidas de similaridade para dados nominais

Num conjunto de dados com variáveis qualitativas, em particular, com variáveis nominais, comummente usam-se medidas de similaridade designadas por coeficientes de semelhança. Estas medidas tomam valores pertencentes ao intervalo [0,1] ou, se expressas em percentagem, no intervalo [0, 100]. Assim, dois objetos, i e j têm o valor do coeficiente de semelhança ou similaridade s_{ij} igual a uma unidade se ambos os objetos tiverem os mesmos atributos para todas as variáveis. Um valor de similaridade zero, indica que os dois indivíduos diferem maximamente para todas variáveis (Everitt, 2011).

Medidas de similaridade para dados nominais com dois níveis (dados binários)

No caso particular dos dados nominais corresponderem a dados binários, a proximidade entre dois indivíduos i e j pode ser medida usando uma tabela de contingência (Tabela 3). Na escolha do coeficiente de semelhança deverá ter-se em conta a utilidade da informação que o respetivo valor fornece ao estudo. Por exemplo, quando a presença comum

de uma característica nos dois objetos em estudo é considerada informativa, é normalmente usado o coeficiente de concordância (Everitt, 2011).

	Indivíduo i			**Total**
	Resultados	**1**	**0**	
Indivíduo j	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d$

Tabela 3: Tabela de contingência de resultados binários para dois indivíduos

Tendo em conta a informação descrita na Tabela 3, a Tabela 4 apresenta uma lista de coeficientes de semelhança (Everitt, 2011; Gower *et al*, 1986). Será usado o símbolo s_{ij} para representar qualquer coeficiente de semelhança entre os objetos i e j e cada coeficiente é uma função diferente dos valores de a, b, c e d identificados na Tabela 3:

Coeficiente	**Fórmula**
Coeficiente de emparelhamento	$s_{ij} = (a+d)/(a+b+c+d)$
Jaccard	$s_{ij} = a/(a+b+c)$
Rogers e Tanimoto	$s_{ij} = (a+d)/[a+2(b+c)+d]$
Sneath e Sokal	$s_{ij} = a/[a+2(b+c)]$
Gower e Lagendre I	$s_{ij} = (a+d)/[a+1/2(b+c)+d]$
Gower e Lagendre II	$s_{ij} = a/[a+1/2(b+c)]$

Tabela 4: Coeficientes de semelhança mais usados para dados binários

2.1.2.3. Medidas de proximidades entre variáveis

No agrupamento de variáveis, a medida de proximidade a usar é uma medida de associação, que corresponde a uma medida de similaridade. Assim, quanto maior for o valor observado para uma medida de associação maior será a proximidade entre variáveis. Para medir a dissimilaridade entre duas variáveis x_i e x_j, registadas de forma emparelhada sobre n indivíduos, calcula-se uma medida de associação (similaridade) entre as duas variáveis e,

seguidamente, por conversão de similaridade para dissimilaridade, é obtida uma medida de dissimilaridade. A seguir listam-se algumas das medidas de associação (i.e., de similaridade) mais usadas para variáveis quantitativas x_i e x_j, onde x_{ki} é o valor da variável da coluna i para o indivíduo k, $k = 1,2, \dots, n$, da matriz de dados X.

- Medida de correlação linear de Pearson

$$s_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)\,(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2 \left(x_{kj} - \overline{x}_j\right)^2}}$$

onde $\overline{x}_i$ representa a média da variável i calculada sobre os n indivíduos.

A medida de correlação de Pearson varia entre [-1, 1], em que:

- ✓ $s_{ij} = 1$ indica uma correlação linear perfeitamente positiva
- ✓ $s_{ij} = -1$ indica uma correlação linear perfeitamente negativa
- ✓ $s_{ij} = 0$ indica que não existe correlação linear.

- Medida do Cosseno

$$s_{ij} = \cos(\alpha) = \frac{\sum_{k=1}^{n} x_{ki} x_{kj}}{\sqrt{\sum_{k=1}^{n}(x_{ki})^2 \left(x_{kj}\right)^2}}$$

Este coeficiente define-se no intervalo [-1, 1] em que:

- ✓ $s_{ij} = 1$ indica que os dois objetos i e j são semelhantes ($\alpha = 0^0$)
- ✓ $s_{ij} = -1$ indica que os dois objetos i e j são em sentidos opostos ($\alpha = 180^0$)
- ✓ $s_{ij} = 0$ indica que os dois objetos i e j são ortogonais ($\alpha = 90^0$)

De seguida, alistam-se medidas de proximidade para variáveis qualitativas binárias com base na Tabela 3, de acordo com Anderberg (1973, pp.84-85):

- baseada na medida de correlação de Pearson sobre dados binários 0-1:

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)}}$$

- baseada na medida do cosseno sobre dados binários 0-1:

$$\frac{a}{\sqrt{(a+b)(a+c)}}$$

A seguir listam-se alguns coeficientes de semelhança mais usados para medir o nível de associação entre duas variáveis qualitativas nominais, os quais são função do valor observado da estatística de qui-quadrado de Pearson χ^2 associada à tabela de contingência das variáveis nominais em avaliação.

- Coeficiente de contingência quadrático

$$\Phi^2 = \frac{\chi^2}{n}$$

- Coeficiente de contingência de Pearson

$$P = \left[\frac{\Phi^2}{1+\Phi^2}\right]^{\frac{1}{2}}$$

- Coeficiente de contingência de Tschuprow

$$T = \left[\frac{\Phi^2}{\sqrt{(r-1)(s-1)}}\right]^{\frac{1}{2}}$$

- Coeficiente V de Cramer

$$C = \left[\frac{\Phi^2}{min(r-1, s-1)}\right]^{\frac{1}{2}}$$

onde r representa o número de linhas e s representa o número da colunas da matriz de dados.

2.2. Métodos hierárquicos

2.2.1. Definição

Os métodos hierárquicos são técnica simples em que os dados na partição sucessivamente, resultando em uma representação hierárquica dos agrupamentos (Everitt, 2011). Portanto essa representação ilustra a visualização dos agrupamentos em cada grupo onde ocorre e com o grau de semelhança entre eles.

Os métodos hierárquicos são subdivididos em métodos aglomerativos e métodos divisivos.

2.2.2. Métodos aglomerativos

Os métodos aglomerativos são os mais utilizados no conjunto dos métodos hierárquicos. Nos métodos aglomerativos os dados são inicialmente distribuídos de modo que cada indivíduo (ou variável) represente um *cluster* e, seguidamente, esses *clusters* são recursivamente agrupados, considerando alguma medida de similaridade e algum critério de aglomeração de *clusters*, até que todos os indivíduos pertençam apenas a um único c*luster*. Notemos que para além de ser necessário definir a medida de proximidade entre dois elementos a considerar, é necessário definir como medir a proximidade entre dois c*lusters* (isto é, grupos de indivíduos), ou seja, definir o critério de agregação entre grupos (Everitt, 2011; Reis, 2001; Sharma, 1996; Timm, 2002). Mais ainda, independentemente da medida de proximidade ou do critério de agregação considerados, no processo de aglomeração, correspondente à selecção dos dois grupos a ser aglomerados, os dois grupos iniciais são obtidos por aqueles que apresentam a menor distância.

Assim, os métodos aglomerativos obedecem ao seguinte algoritmo para agrupar n objetos, itens ou indivíduos:

a. O processo inicia com n *clusters* ou grupos, cada um com um objeto, e a respetiva matriz de distâncias ou similaridades através de alguma medida de proximidade;
b. Identifica-se na matriz o par i e j com a menor distância ou maior similaridade, sendo a distância entre i e j dada por d_{ij} ;
c. Os objetos i e j são agrupados em um c*luster* que passa a ser denominado ij e a matriz de distâncias é atualizada:
    - Eliminando-se a linha e a coluna referentes aos objetos i e j;
    - Calculando-se as distâncias entre os demais objetos e o grupo ij
d. Repetir os itens (b) a (c) até que todos os objetos formem um único c*luster*.

Os passos acima aplicam-se a todos os critérios de agregação, nomeadamente o do vizinho mais próximo (s*ingle linkage*), do vizinho mais afastado (*complete linkage*), o critério de média (*average*), o critério do centróide e o critério de *Ward.*

O processo de aglomeração dos objetos ou variáveis pode ser ilustrado através de um dendrograma. Um dendrograma é uma representação gráfica onde se visualizam os passos realizados na aglomeração numa Análise de *Clusters*, mostrando como os grupos se vão formando, que permite dar uma ideia dos valores das medidas de proximidade dos grupos dentro de cada etapa no processo aglomerativo (Everitt, 2011; R. Johnson *et al*, 2014).

2.2.2.1. Método do vizinho mais próximo (s*ingle linkage*)

O método do vizinho mais próximo é um método que utiliza a distância entre dois itens mais próximos (vizinhos) como a distância mínimo entre dois grupos. Por exemplo, segundo este método, a distancia entre os dois grupos (i, j) e (k) é dada por

$$d_{(ij)k} = min(d_{ik}, d_{jk})$$

Exemplo 4: Consideremos a matriz de dados quantitativos com 5 indivíduos (amostras) e 6 variáveis (V1, …, V6) apresentada na Tabela 5.

Objeto	V1	V2	V3	V4	V5	V6
Amostra 1	1	2	3	4	5	6
Amostra 2	5	4	1	8	7	9
Amostra 3	6	5	4	2	7	9
Amostra 4	6	4	2	1	3	7
Amostra 5	9	2	1	4	7	8

Tabela 5: Matriz de dados

Para a obtenção de grupos de amostras usando o critério do vizinho mais próximo, poderíamos usar diversas distâncias, mas, no contexto deste exemplo, usaremos a distância Euclidiana para calcular a proximidade (dissimilaridade) entre amostras. Para obter a matriz de dissimilaridades através da distância Euclidiana, começamos por calcular a distância entre a amostra 1 e a amostra 2, dada por:

$$d_{12} = \sqrt{(1-5)^2 + (2-4)^2 + (3-1)^2 + (4-8)^2 + (5-7)^2 + (6-9)^2} \approx 7{,}3$$

a seguir, a distância entre a amostra 1 e a amostra 3, dada por:

$$d_{13} = \sqrt{(1-6)^2 + (2-5)^2 + (3-4)^2 + (4-2)^2 + (5-7)^2 + (6-9)^2} \approx 7{,}2$$

E assim sucessivamente, da mesma maneira para obter as distâncias entre as outras amostras. No fim, obtém-se a matriz de dissimilaridades:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & 5{,}1 & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Procedemos à Análise de *Clusters* aplicando o seu algoritmo partindo do facto de que cada objeto é um cluster e da matriz de dissimilaridades (distâncias). Agora, identifica-se a menor distância (maior similaridade), a qual é dada pela distância entre a amostra 3 e a amostra 4 com valor $d_{43} = 5{,}1$; assim, o novo *cluster* é : (43).

- Calcular novas distâncias:

$$d_{(34)1} = \min\{d_{(31)}, d_{(41)}\} = \min(7{,}2\,;\,6{,}6) = 6{,}6$$

$$d_{(43)2} = \min\{d_{(42)}, d_{(32)}\} = \min(8{,}4\,;\,6{,}9) = 6{,}9$$

$$d_{(43)5} = \min\{d_{(45)}, d_{(35)}\} = \min(6{,}3\,;\,5{,}7) = 5{,}7$$

- A nova matriz de dissimilaridade é

$$\begin{array}{c} \\ (43) \\ 1 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (43) & 1 & 2 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 6{,}6 & 0 & & \\ 6{,}9 & 7{,}3 & 0 & \\ (5{,}7) & 8{,}3 & 6{,}1 & 0 \end{bmatrix} \end{array}$$

Procedendo da mesma maneira, escolhe-se o valor mínimo de nova matriz recalculada, obtendo um novo *cluster* (345), e assim sucessivamente até encontrar uma única matriz de dados com um *cluster* com cinco objetos.

Passo	Distância	Grupos
1	$d_{(43)} = 5{,}1$	(3,4) (1) (2) (5)
2	$d_{(34)5} = 5{,}7$	(3,4,5) (1) (2)
3	$d_{(435)2} = 6{,}1$	(3,4,5,2) (1)
4	$d_{(4352)1} = 6{,}6$	(1,2,3,4,5)

Os agrupamentos e os níveis de distância em que ocorrem estão claramente ilustrados no dendrograma (Figura 2).

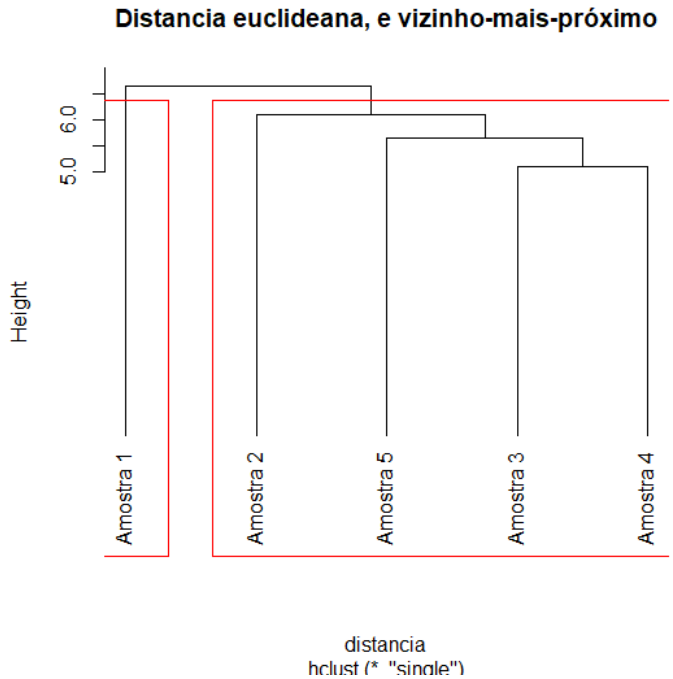

Figura 2:Dendrograma do critério do vizinho mais próximo do Exemplo 4

2.2.2.2. Método do vizinho mais distante (*complete linkage*)

Neste método a distância entre dois grupos é definida como sendo a distância entre os seus elementos mais afastados. Assim, por exemplo, a distância entre os grupos (i, j) e (k) é dada por

$$d_{(ij)k} = max(d_{ik}, d_{jk})$$

Exemplo 5: Relativamente à matriz de distâncias do Exemplo 4, dada por

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & (5{,}1) & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

identifica-se a menor distância (maior similaridade) dada por $d_{(34)} = 5{,}1$ e obtenção então do novo *cluster* (34)

- Calcular novas distâncias

$$d_{(34)1} = \max\{d_{(31)}, d_{(41)}\} = \max(7{,}2;\ 6{,}6) = 7{,}2$$

$$d_{(34)2} = \max\{d_{(32)}, d_{(42)}\} = \max(6{,}9;\ 8{,}4) = 8{,}4$$

$$d_{(34)5} = \max\{d_{(35)}, d_{(45)}\} = \max(5{,}7; 6{,}3) = 6{,}3$$

- Nova matriz de dissimilaridade é

$$\begin{array}{c} \\ (43) \\ 1 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (43) & 1 & 2 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 7{,}2 & 0 & & \\ 8{,}4 & 7{,}3 & 0 & \\ 6{,}3 & 8{,}3 & (6{,}1) & 0 \end{bmatrix} \end{array}$$

Procedendo de modo similar, escolhendo o valor mínimo desta nova matriz identifica-se o novo cluster (25) e assim sucessivamente até encontrar um *cluster* com cinco objetos. Todo este processo de agrupamento pode ser resumido no quadro seguinte:

Passo	Distâncias	Grupos
1	$d_{(43)} = 5{,}1$	(3,4) (1) (2) (5)
2	$d_{(52)} = 6{,}1$	(3,4) (52) (1)
3	$d_{(34)1} = 7{,}2$	(3,4,1) (52)
4	$d_{(341)(52)} = 8{,}7$	(1,2,3,4,5)

Assim, apresentado no dendrograma em seguida:

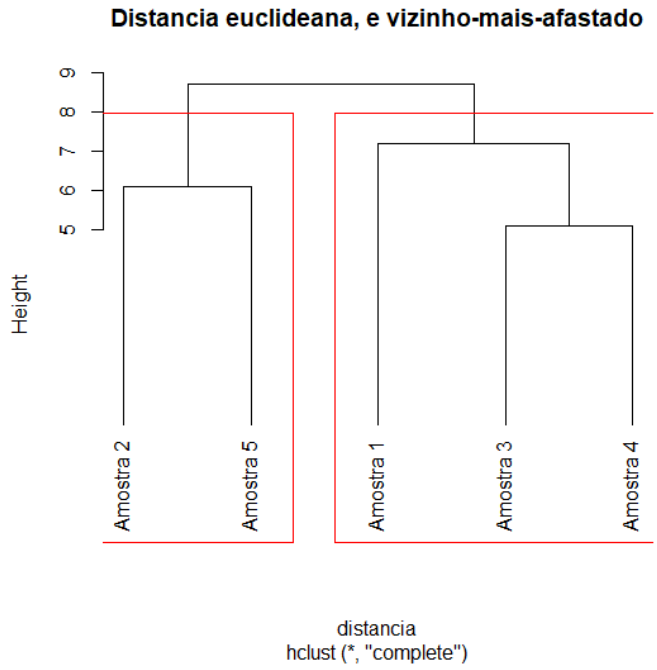

Figura 3:Dendrograma usando o método do vizinho mais afastado do Exemplo 5

2.2.2.3. Método de média (*average linkage*)

No método de ligação média, a distância entre dois *clusters* é obtida tomando-se a distância média entre todos os pares possíveis de objetos pertencentes aos dois *clusters* (Sharma, 1996). Por exemplo, a distância entre dois grupos A e B é dada por:

$$d_{AB} = \frac{1}{N}\sum_{i}\sum_{j} d_{ij}$$

onde N é o número de pares de elementos entre os dois *clusters* A e B e d_{ij} representa a distância entre o elemento i do *cluster* A e o elemento j do *cluster* B . Esta estratégia é intermédia relativamente às duas descritas anteriormente.

Exemplo 6: Relativamente a matriz de distâncias do Exemplo 4 representada por:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & (5{,}1) & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

o primeiro agrupamento é formado pelo conjunto dos sujeitos amostra 3 e amostra 4. A distância entre o *cluster* (34) e o sujeito amostra 1 é a distância

$$d_{(34)1} = \frac{d_{31} + d_{41}}{2} = \frac{7{,}2 + 6{,}6}{2} = 6{,}9$$

A matriz de semelhança resultante após a formação do primeiro grupo é dada por:

$$D_1 = \begin{array}{c} \\ (34) \\ 1 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (34) & 1 & 2 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 6{,}9 & 0 & & \\ 7{,}65 & 7{,}3 & 0 & \\ (6) & 8{,}7 & 6{,}1 & 0 \end{bmatrix} \end{array}$$

Para restantes passos, procede-se da mesma maneira escolhendo o valor mínimo da nova matriz de distâncias recalculada, e assim sucessivamente até encontrar uma única matriz de dados com cinco objetos. Neste caso, obtém-se:

Passo	**Distâncias**	**Clusters**
1	$d_{(34)} = 5{,}1$	(34), 1, 2, 5
2	$d_{(345)} = 6$	(345), 1, 2
3	$d_{(3452)} = 6{,}5$	(3452), 1
4	$d_{(34521)} = 7{,}75$	(1,2,3,4,5)

Dendrograma do método *Average*:

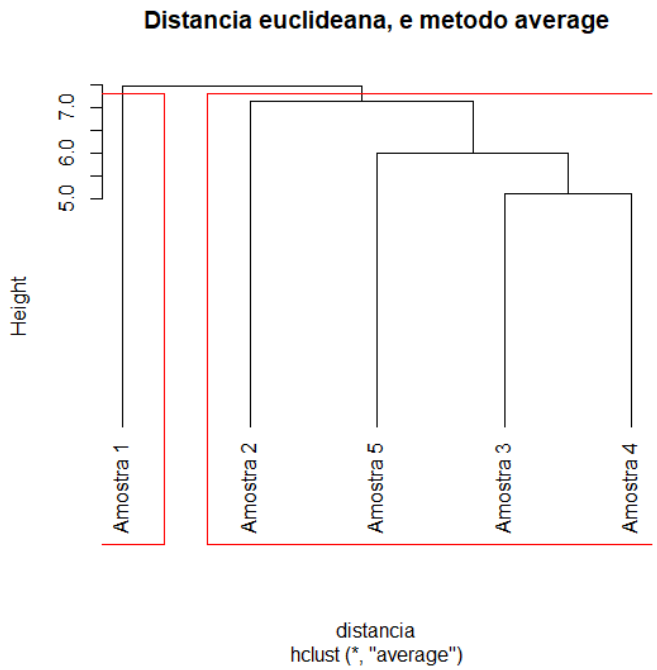

Figura 4:Dendrograma usando o método Average do Exemplo 6

2.2.2.4. Método centróide

O método centróide calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis. Concretamente, dados dois grupos A e B, a distância entre eles, é igual à distância entre os seus centróides , $\bar{x}_A, \bar{x}_B$, isto é:

$$d_{AB} = d(\bar{x}_A, \bar{x}_B)$$

onde:

$$\bar{x}_A = \frac{1}{n_A}\sum_{i\in A} x_i \text{ e } \bar{x}_B = \frac{1}{n_B}\sum_{i\in B} x_i$$

x_ié um vetor das p observações do objeto i; $\bar{x}_A$é um vector formado pelas médias aritméticas das p variáveis, calculadas para os n_A objetos que pertencem ao grupo A. Analogamente para $\bar{x}_B$ e aos n_B objetos que pertencem ao grupo B.

No critério de agregação centróide também pode ser aplicado a fórmula de recorrência de Lance e Williams (Wunsch II et al, 2008). Concretamente, a distância entre um grupo C e um grupo (A, B) formado pela fusão de dois grupos A e B, é dada por :

$$d_{(AB)C} = \alpha_A d_{AC} + \alpha_B d_{BC} + \beta d_{AB} + \gamma[d_{AC} - d_{AB}]$$

em que $\alpha_A, \alpha_B, \beta, \gamma$ são parâmetros que ou são constantes ou dependem do número de objetos em cada grupo n_A, n_B, e n_C e d_{AB} é a dissimilaridade entre os grupos A e B. Para o método centróide, tem-se:

$$\alpha_A = \frac{n_A}{n_A+n_B}, \alpha_B = \frac{n_B}{n_A+n_B}, \ \beta = \frac{-n_A n_B}{(n_A+n_B)^2} \text{ e } \gamma = 0$$

Exemplo 7: Considere-se a matriz distância do Exemplo 4 representada por:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & (5{,}1) & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Passo 1: O primeiro agrupamento é formado pelo conjunto dos sujeitos amostra3 e amostra 4, com a distância 5,1. A seguir é calculada a distância entre este novo *cluster* e cada um dos restantes objetos:

$$d_{(34)1} = \frac{1}{2}(d_{13} + d_{14}) - \frac{1}{4}d_{34} = \frac{1}{2} \times (7{,}2 + 6{,}6) - \frac{1}{4} \times 5{,}1 = 5{,}6$$

$$d_{(34)2} = \frac{1}{2}(d_{23} + d_{24}) - \frac{1}{4}d_{34} = \frac{1}{2} \times (6{,}9 + 8{,}4) - \frac{1}{4} \times 5{,}1 = 6{,}37$$

$$d_{(34)5} = \frac{1}{2}(d_{35} + d_{45}) - \frac{1}{4}d_{34} = \frac{1}{2} \times (5{,}7 + 6{,}3) - \frac{1}{4} \times 5{,}1 = 4{,}7$$

Reescrever novamente a matriz de similaridade:

$$\begin{array}{c} \\ 34 \\ 1 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 34 & 1 & 2 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 5{,}6 & 0 & & \\ 6{,}37 & 7{,}3 & 0 & \\ (4{,}7) & 8{,}7 & 6{,}1 & 0 \end{bmatrix} \end{array}$$

Para restantes passos, procede-se da mesma maneira (de escolher o valor mínimo de nova matriz distância que se recalculou); por fim recalcular novamente até encontrar uma única matriz de dados com cinco objetos. Neste caso, tem-se:

Passos	**Distância**	**Grupos formados**
1º	5,1	(34), (1), (2), (5)
2º	4,7	(345), (2), (1)
3º	5,06	(2345), (1)
4º	5,34	(12345)

Dendrograma do método de centróide:

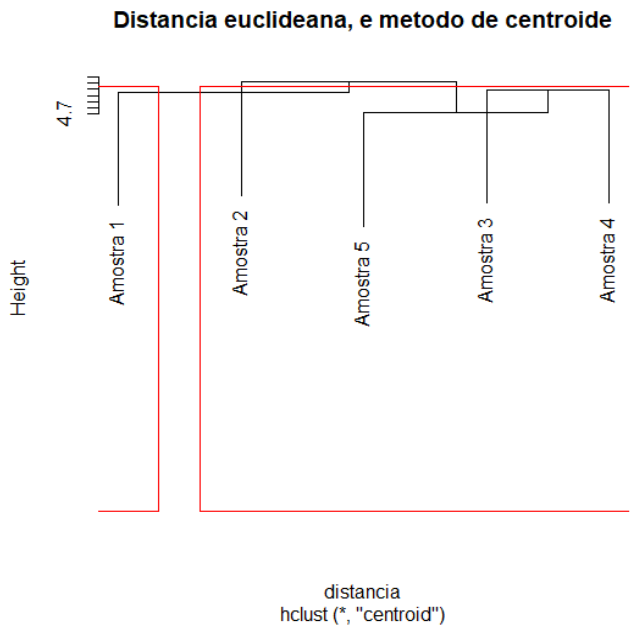

Figura 5:Dendrograma usando o método de centróide do Exemplo 7

2.2.2.5. Método de Ward

No método de Ward o objetivo é a formação de *clusters* por forma a maximizar a homogeneidade dentro de cada *cluster*. A soma de quadrados dentro de um *cluster*, dada pela soma de quadrados dos desvios à média , é a medida de homogeneidade usada (Sharma, 1996). O *cluster* obtido em cada etapa é o *cluster* com menor valor para a soma de quadrados dos erros (SQE). Para um grupo qualquer A, a SQE do grupo é dada por

$$SQE_A = \sum_{i=1}^{n_A} \sum_{j=1}^{p} (x_{ijA} - \bar{x}_{jA})^2 \qquad (2.1)$$

Para obter a soma de quadrados dos desvios quando se combinam (i.e., fundem) dois *clusters* A, B para formar um novo *cluster* C no próximo nível, pode recorrer-se ao incremento na SQE, dada por:

$$\Delta SQE = SQE_C - (SQE_A + SQE_B)$$

onde a soma de quadrados dos erros do novo *Cluster* C é dada por:

$$SQE_C = \sum_{i=1}^{n_C} \sum_{j=1}^{p} (x_{ijC} - \bar{x}_{jC})^2$$

Este incremento ΔSQE é geralmente usado como uma medida de proximidade entre grupos, sendo o objetivo minimizar este incremento quando os dois grupos são unidos. Neste incremento a soma dos quadrados dos desvios dentro do novo *cluster* C, SQE_C, é dada por (Kaufman *et al*, 2005):

$$\frac{n_A n_B d_{AB}^2}{n_A + n_B}$$

em que:

$$d_{AB}^2 = \sum_{j=1}^{p} (\bar{x}_{jA} - \bar{x}_{jB})^2 \qquad (2.2)$$

A medida da distância total entre A e B é dada por $n_A n_B d_{AB}^2$. Uma vez que existem $n_A + n_B$ objetos, então

$$\frac{n_A n_B d_{AB}^2}{n_A + n_B} \tag{2.3}$$

representa uma medida de distância média e é equivalente à alteração na soma dos quadrados dos erros dentro do grupo, ou seja à soma de quadrados dos erros incremental que resulta da combinação dos grupos A e B (Rencher, 2012).

Exemplo 8: Consideremos o conjunto de dados do Exemplo 4 para aplicar o método de Ward usando duas variantes no cálculo das dissimilaridades entre *clusters*, uma baseada na SQE (equação 2.1) e outra baseada no ΔSQE (a qual é calculada usando a equação 2.3). Para ilustração detalhamos o cálculo de SQE e ΔSQE quando resultam:

Caso 1: quatro *clusters* possíveis (Amostra 1, Amostra 2), Amostra 3, Amostra 4 e Amostra 5

Caso 2: três *clusters* possíveis (Amostra 1, Amostra 3, Amostra 4), Amostra 2 e Amostra 5. Nestas circunstâncias, tem-se:

- O valor de SQE para o Caso 1, é dado por:

$$SQE_{\{(1,2),3,4,5\}} = (1-3)^2 + (2-3)^2 + (3-2)^2 + (4-6)^2 + (5-6)^2 + (6-7{,}5)^2 + (5-3)^2 + (4-3)^2 + (1-2)^2 + (8-6)^2 + (7-6)^2 + (9-7{,}5)^2 + 0 + 0 + 0 = 26{,}5$$

- O valor de SQE para o Caso 2, é dado por:

$$SQE_{\{(1,3,4),2,5\}} = (1-4{,}3)^2 + (2-3{,}6)^2 + (3-3)^2 + (4-2{,}3)^2 + (5-5)^2 + (6-7{,}3)^2 + (6-4{,}3)^2 + (5-3{,}6)^2 + (4-3)^2 + (2-2{,}3)^2 + (7-5)^2 + (9-7{,}3)^2 + (6-4{,}3)^2 + (4-3{,}6)^2 + (2-3)^2 + (1-2{,}3)^2 + (3-5)^2 + (7-7{,}3)^2 + 0 + 0 = 40{,}69$$

- O valor de ΔSQE para o Caso 1, é dado por:

$$\Delta SQE_{\{(1,2),3,4,5\}} = \frac{n_{(1)} n_{(2)} d_{(1)(2)}^2}{n_{(1)} + n_{(2)}} = \frac{53}{2} = 26{,}5$$

onde o valor de $d^2_{(1)(2)}$ foi calculado utilizando a equação (2.2), sendo o centróide $\bar{x}_1$ dado pela própria Amostra 1 e o centróide $\bar{x}_2$ dado pela própria Amostra 2; ou seja:

$$d^2_{(1)(2)} = (1-5)^2 + (2-4)^2 + (3-1)^2 + (4-8)^2 + (5-7)^2 + (6-9)^2 = 53$$

- O valor de ΔSQE para o Caso 2 é dado por:

$$\Delta SQE_{\{(1,3,4),2,5\}} = \frac{n_{(1)} n_{(34)} d^2_{(1)(34)}}{n_{(1)} + n_{(34)}} = \frac{41{,}5}{2} = 20{,}75$$

onde o valor de $d^2_{(1)(34)}$ foi calculado utilizando a equação (2.2), sendo o centróide $\bar{x}_1$ dado pela própria Amostra 1 e o centróide $\bar{x}_{(34)} = (6; 4{,}5; 3; 1{,}5; 5; 8)$; ou seja:

$$d^2_{(1)(34)} = (1-6)^2 + (2-4{,}5)^2 + (3-3)^2 + (4-1{,}5)^2 + (5-5)^2 + (6-8)^2 = 41{,}5$$

Assim, procedendo de forma similar para os restantes *clusters*, obtiveram-se os seguintes resultados para SQE e ΔSQE:

Soluções de clusters	Número de clusters				SQE	ΔSQE
	1	**2**	**3**	**4**		
(a) Todas as soluções com quatro *clusters* possíveis						
	(1,2)	3	4	5	26,5	26,5
	(1,3)	2	4	5	26	26
	(1,4)	2	3	5	22	22
	(1,5)	2	3	4	38	38
	1	(2,3)	4	5	25,5	25,5
	1	(2,4)	3	5	35,5	35,5
	1	(2,5)	3	4	66,5	66,5
	1	2	**(3,4)**	5	**13**	**13**
	1	2	(3,5)	4	16	16
	1	2	3	(4,5)	20	20
(b) Todas as soluções com três *clusters* possíveis						
	(1,3,4)	2	5		40,69	20,75
	(2,3,4)	1	5		48,05	26,25
	(3,4,5)	1	2		**32,7**	**14,75**
	(3,4)	(1,2)	5		39,5	39,5
	(3,4)	(1,5)	2		51	51
	(3,4)	(2,5)	1		79,5	79,5
(c) Todas as soluções com dois *clusters* possíveis						
	(3,4,5)	(1,2)			59,19	41,25
	(3,4)	**(1,2,5)**			**55,33**	**34,525**
(d) Todas as soluções com um único *cluster* possível						
	(1,2,3,4,5)				**97**	**15,73**

Uma outra forma de obter a distância de um novo cluster é usando a fórmula de recorrência de Lance-Williams, a qual dá a distância entre um grupo C e um grupo A e B, formado pela fusão de dois grupos A e B, como sendo igual a:

$$d_{(AB)C} = \alpha_A d_{AC} + \alpha_B d_{BC} + \beta d_{AB} + \gamma[d_{AC} - d_{AB}]$$

em que α_A, α_B, β e γ são parâmetros que ou são constantes ou dependem do número de objetos em cada grupo n_A, n_B, n_C, e d_{AB} é a dissimilaridade entre grupo A e B. Para o método de Ward. o critério de Lance-Williams é

$$\alpha_A = \frac{n_A+n_C}{n_A+n_B+n_C}, \alpha_B = \frac{n_B+n_C}{n_A+n_B+n_C}, \beta = \frac{-n_C}{n_A+n_B+n_C} \text{ e } \gamma = 0$$

No pacote `stats`, do software estatístico R, existem duas variantes do método Ward: `Ward.D` e `Ward.D2`. A diferença entre as duas variantes está na medida de distância. Na variante `Ward.D` é considerada a dissimilaridade das entradas em termos das distâncias euclidianas (não quadrado), enquanto que no método `Ward.D2` esta está em termos das distâncias euclidianas ao quadrado. As distâncias euclidianas ao quadrado são as que nos interessam uma vez que surgem no critério de soma de quadrados dos erros (Murtagh and Legendre ,2014).

Exemplo 9: Consideremos novamente a matriz de distâncias do Exemplo 4, dada por

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & (5{,}1) & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Com essa matriz de dissimilaridade, utilizando a fórmula de recorrência de Lance-Williams `Ward.D`, o método de Ward segue os seguintes passos:

Passo 1: identificar-se na matriz de dissimilaridade os objetos mais similares: amostra 3 e amostra 4, com distância igual a 5.1, formando o grupo inicial.

- Calcular as distâncias Euclidianas para entre este novo *cluster* e cada uma das restantes amostras, que faz a combinação do grupo inicial (34)

$$d_{(34)1} = \frac{2}{3}(d_{13} + d_{14}) - \frac{1}{3}d_{34} = 9{,}2 - 1{,}7 = 7{,}5$$

$$d_{(34)2} = \frac{2}{3}(d_{23} + d_{24}) - \frac{1}{3}d_{34} = 10{,}2 - 1{,}7 = 8{,}5$$

$$d_{(34)5} = \frac{2}{3}(d_{35} + d_{45}) - \frac{1}{3}d_{34} = 8 - 1{,}7 = 6{,}3$$

- Reescrever a nova matriz de dissimilaridade:

$$\begin{array}{c} \\ 34 \\ 1 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 34 & 1 & 2 & 5 \end{array} \\ \begin{bmatrix} 0 & & & \\ 7{,}5 & 0 & & \\ 8{,}5 & 7{,}3 & 0 & \\ 6{,}3 & 8{,}7 & (6{,}1) & 0 \end{bmatrix} \end{array}$$

Para outros passos, repetindo o procedimento de escolher o valor mínimo de nova matriz de dissimilaridade, até encontrar uma única matriz de distâncias com os cinco objetos agrupados. Obtém-se:

Passos	SQE	Grupos formados
1º	5,1	(34), (1), (2), (5)
2º	6,1	(52), (34), (1)
3º	7,5	(341), (25)
4º	8,53	(12345)

Os resultados foram apresentados no dendrograma da Figura 6.

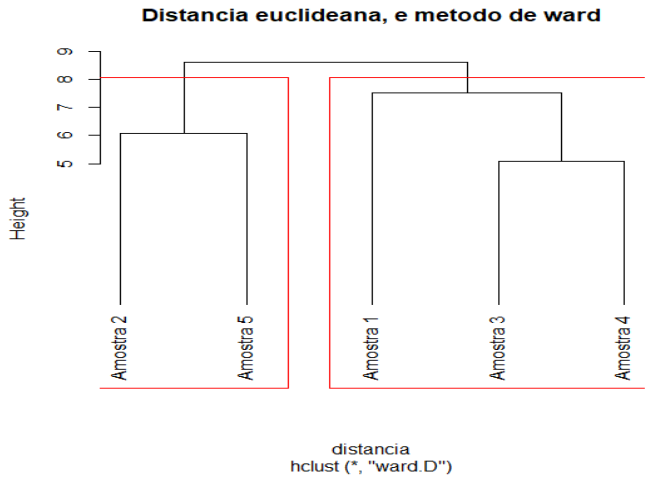

Figura 6: Dendrograma obtido pelo método de Ward do Exemplo 9.

Em forma de resumo, importa referir que as medidas de distância intra-grupos usadas por algumas técnicas hierárquicas aglomerativas usuais de construção de *cluster*, são obtidas por escolha adequada dos parâmetros $\alpha_A, \alpha_B, \beta, \gamma$ na fórmula de recorrência de Lance-Williams, de acordo com a Tabela 6.

Critério de ligação	**Parâmetros de Lance-Williams**			
	α_A	α_B	β	γ
Ligação simples	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Ligação completa	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ligação média (Average)	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	0	0
Método centróide	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	$\frac{-n_A n_B}{(n_A+n_B)^2}$	0
Método mediana	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Método de Ward	$\frac{n_A+n_C}{n_A+n_B+n_C}$	$\frac{n_B+n_C}{n_A+n_B+n_C}$	$\frac{-n_C}{n_A+n_B+n_C}$	0

Tabela 6: Parâmetros de Lance-Williams para vários métodos hierárquicos.

2.2.3. Métodos divisivos

Os métodos divisivos apresentam um procedimento inverso aos métodos aglomerativos (Figura 7): parte de um grande *cluster* e, por passos sucessivos de divisão de subgrupos (*clusters*), estabelece novos subgrupos parando quando é obtido um elemento em cada *cluster*. A divisão de dois subgrupos distintos é determinada em função de algum critério de dissimilaridade (distância) (Berkhin, 2002).

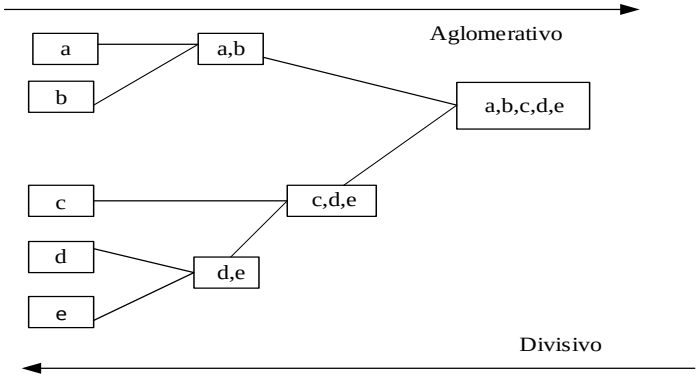

Figura 7: Gráfico ilustrando a relação de procedimento entre um método aglomerativo e um método divisivo.

Embora seja menos utilizado do que métodos aglomerativos, os métodos divisivos têm a vantagem que a maioria dos usuários está interessado na estrutura principal dos seus dados, o que é revelado desde o início de um método divisivo.

Computacionalmente os métodos divisivos são geralmente exigentes. No caso particular das variáveis serem binárias, os métodos divisivos, conhecidos por métodos monotéticos, são computacionalmente mais eficientes.

2.2.3.1. Métodos divisivos monotéticos

Nos métodos divisivos monotéticos todas as variáveis são binárias. Uma forma de definir os *clusters*, nos métodos divisivos monotéticos, é identificar a variável que globalmente se encontra mais associada com as restantes variáveis em cada fase de formação dos clusters. A forma de identificar tal variável corresponde a somar todas as similaridades dessa variável com as restantes tendo em conta os indivíduos do cluster em divisão. A variável com a maior soma será a variável escolhida para a separação do cluster nesse passo. Os indivíduos com valores iguais a 1 nessa variável escolhida formarão um cluster e os restantes o outro cluster. Porque a separação é feita com base numa única variável binária, este método divisivo chama-se monotético. A medida de similaridade usada é a chamada associação,

especialmente utilizada em Ecologia (Williams e Lambert, 1959). Por exemplo, para um par de variáveis binárias, v_i e v_j com valores 0 e 1, as frequências observadas podem ser:

v_i	v_j	
	1	0
1	a	b
0	c	d

Tabela 7: Tabela contingência num par de variáveis binárias nos métodos monotéticos.

As medidas de associação mais comuns (somadas para todos os pares de variáveis) são as seguintes:

- $|ad-bc|$ (2.4)
- $(ad-bc)^2$
- $\frac{(ad-bc)^2 n}{[(a+b)(a+c)(b+d)(c+d)]}$
- $\sqrt{\frac{(ad-bc)^2 n}{[(a+b)(a+c)(b+d)(c+d)]}}$
- $\frac{(ad-bc)^2}{[(a+b)(a+c)(b+d)(c+d)]}$

As duas primeiras medidas acima têm a vantagem de não produzir problema a nível computacional, se algum total marginal da tabela de contingência for zero. As três últimas acima listadas estão relacionadas com a estatística de qui-quadrado de Pearson χ^2 (nomeadamente, com o coeficiente de contingência quadrático).

Exemplo 10: Considere-se a seguinte matriz de dados binários, com cinco indivíduos e três variáveis.

indivíduo	x_1	x_2	x_3
1	0	1	0
2	1	1	1
3	1	1	0
4	1	1	1
5	0	0	0

Para aplicar o método monotético, sobre as variáveis $\{x_1, x_2, x_3\}$ e usando como critério de homogeneidade a medida de associação (2.4), devemos calcular o valor da associação de todos os pares de variáveis: $\{x_1, x_2\}$, $\{x_2, x_3\}$, $\{x_1, x_3\}$ a fim de identificar a primeira variável binária que efetuará a separação do cluster inicial (1,2,3,4,5). Por exemplo, para o par $\{x_2, x_3\}$, constrói-se a respetiva tabela contingência:

		$variável\ 2$		
		1	0	Total
$variável\ 3$	1	2	1	3
	0	1	1	2
	Total	3	2	5

onde, pela equação (2.4), obtém-se

$$|ad - bc| = |2 - 1| = 1.$$

Usando o mesmo procedimento para os restantes pares, obtém-se: para o par $\{x_1, x_2\}$, o valor de associação igual a 3, e para o par $\{x_1, x_3\}$ o valor de associação igual a 4. Assim, para obter os maiores somas de similaridades entre pares de variáveis calculamos conforme a seguinte tabela:

Variável	**Os pares de variáveis**	**Valor associação**	**Soma de similaridades**
$\boldsymbol{x_1}$	$\{x_1, x_2\}$	3	7
	$\{x_1, x_3\}$	4	
x_2	$\{x_2, x_1\}$	3	4
	$\{x_2, x_3\}$	1	
x_3	$\{x_3, x_1\}$	4	5
	$\{x_3, x_2\}$	1	

A variável com maior soma de similaridades é x_1. Com base nesta o cluster {1,2,3,4,5} fica dividido em {2,3,4} e {1,5}. Procedendo de forma análoga sobre o cluster {2,3,4} obtemos associações nulas, pelo que este cluster não é mais dividido. O mesmo acontece para o cluster {1,5}. Logo, o processo divisivo está concluído.

2.2.3.2. Métodos divisivos politéticos

Os métodos divisivos politéticos são mais parecidos com métodos aglomerativos pois recorrem a uma matriz de proximidade tomando todas as variáveis da matriz de dados. O procedimento de MacNaughton-Smith (Everitt 2011) evita considerar todas as divisões possíveis, um problema potencial dos métodos divisivos politéticos.

Um processo dos métodos divisivos foi introduzido por Kaufman e Rousseeuw (2005), conhecido por "*DIANA*" *(Divisive Analysis Clustering)*, que está implementado no S-plus e no R. No método Diana é calculado um coeficiente, denominado coeficiente divisivo, o qual mede a qualidade do agrupamento dos dados e é dado por:

$$CD = 1 - \frac{1}{n}\sum_{i=1}^{n} \bar{d}_{(i)}$$

onde n é o número total de objetos do conjunto de dados e $\bar{d}_{(i)}$ é o diâmetro (uniformizado entre 0 e 1) do último cluster ao qual o objeto *i* pertencia antes de ser retirado desse cluster, sendo que o diâmetro de um cluster é a maior dissimilaridade entre quaisquer dois elementos desse cluster. Prova-se que o CD varia entre 0 e 1, onde valores baixos do coeficiente corresponde a uma má estrutura do agrupamento, indicando que nenhum agrupamento foi encontrado, e valores próximos de 1 indica que está identificada uma clara estrutura de agrupamento.

A seguir, apresenta-se o algoritmo de *DIANA* :

1. Iniciar com uma matriz de distâncias (dissimilaridade) $D_{n\times n}$
2. Encontrar o objeto, o que tem a maior dissimilaridade média relativamente a todos os outros objetos. Este objeto inicia um novo *cluster* sendo uma espécie de grupo dissidente.
3. Para cada objeto i fora grupo dissidente, calcular a dissimilaridade média.
4. Calcular a diferença de distâncias médias

$$D_i = \left[Media\ d_{(ij)};\ j \notin R_{grupo\ dissidente}\right] - \left[Media\ d_{(ij)};\ j \in R_{grupo\ dissidente}\right]$$

5. Encontrar um objeto h para o qual a diferença D_h é maior. Se D_h é positivo, então o objeto h está mais longe do grupo dissidente.
6. Repetir os passos 2 e 3 até que todas as diferenças D_h sejam negativas. Nesta altura são formados dois novos *clusters*.
7. Selecionar o *cluster* com o maior diâmetro. O diâmetro de um *cluster* é o maior dissimilaridade entre quaisquer dois dos seus objetos. Em seguida, dividir este conjunto, seguindo os passos 2 até 5.
8. Repetir a etapa 6 até todos os aglomerados contenham apenas um único objeto.

Exemplo 11: Consideremos a matriz de dissimilaridade relativa ao Exemplo 4:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & 5{,}1 & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Vamos aplicar o algoritmo *Diana*. Parte-se de um *cluster* formado por todos os indivíduos: (1,2,3,4,5). O diâmetro deste *cluster* é 8.7 (valor máximo de todas as distâncias entre pares de indivíduos. Para definir a divisão deste *cluster*, calcula-se, em primeiro lugar, a dissimilaridade média. Obtém-se:

Indivíduo para grupo dissidente	**Distância média**
1.	$\frac{7{,}3+7{,}2+6{,}6+8{,}7}{4}=7{,}45$
2.	$\frac{7{,}3+6{,}9+8{,}4+6{,}1}{4}=7{,}175$
3.	$\frac{7{,}2+6{,}9+5{,}1+5{,}7}{4}=6{,}225$
4.	$\frac{6{,}6+8{,}4+5{,}1+6{,}3}{4}=6{,}6$
5.	$\frac{8{,}7+6{,}1+5{,}7+6{,}3}{4}=6{,}7$

Verifica-se que o indivíduo 1 possui a maior dissimilaridade, originando os *clusters* iniciais: grupo dissidente (1) e restante grupo (2,3,4,5). Calculando a dissimilaridade média para os indivíduos do restante grupo, a dissimilaridade média entre cada elemento do grupo e o elemento retirado e as diferenças desses valores, obtemos:

Indivíduo para grupo dissidente	**Distância média para o grupo restante (A)**	**Distância média para o grupo dissidente (B)**	**Diferença (A-B)**
2.	$\frac{6,9+8,4+6,1}{3}=7,13$	7,3	-0,17
3.	$\frac{6,9+5,1+5,7}{3}=5,9$	7,2	-1,3
4.	$\frac{8,4+5,1+6,3}{3}=6,6$	6,6	0
5.	$\frac{6,1+5,7+6,3}{4}=6,03$	8,7	-2,67

Na tabela acima, verifica-se que o elemento 4 possui a maior diferença positiva, então é retirado do restante grupo, grupo (2,3,4,5), e agrupado ao dissidente inicial (1).

Recalculando as dissimilaridades médias entre os elementos do restante grupo (2,3,5), a dissimilaridade dos elementos desse grupo em relação ao grupo dissidente (1,4) e a diferença desses valores, obtemos:

Indivíduo para grupo dissidente	**Distância média para o restante grupo (A)**	**Distância média para o grupo dissidente (B)**	**Diferença (A-B)**
2.	$\frac{6,9+6,1}{2}=6,5$	$\frac{7,3+8,4}{2}=7,85$	-0,79
3.	$\frac{6,9+5,7}{2}=6,3$	$\frac{7,2+5,1}{2}=6,15$	0,45
5.	$\frac{6,1+5,7}{2}=5,9$	$\frac{8,7+6,3}{2}=7,5$	-0,3

Na tabela acima, verifica-se que o elemento 3 possui a maior diferença positiva, então é retirado do grupo (2,3,5) e agrupado ao grupo dissidente (1,4).

Recalculando a dissimilaridade média entre os elementos do grupo (2,5), a dissimilaridade dos elementos desse grupo em relação ao grupo dissidente (1,3,4) e a diferença desses valores, obtemos:

Indivíduo para grupo dissidente	**Distância média para o restante grupo (A)**	**Distância média para o grupo dissidente (B)**	**Diferença (A-B)**
2.	6,1	$\frac{7{,}3+6{,}9+8{,}4}{3}=7{,}53$	-1,43
5.	6,1	$\frac{8{,}7+5{,}7+6{,}3}{3}=6{,}9$	-0,8

Como todas as diferenças são agora negativas, significa que a distância de todos os objetos para o restante grupo já é maior do que a distância para o grupo dissidente, pelo que a composição está estável e está então identificada os dois novos *clusters*: grupo restante(1,3,4) e grupo dissidente (2,5). O diâmetro do grupo (1,3,4) é 7,2 (valor máximo de d_{ij} para $i,j = 1{,}3{,}4$ e $i \neq j$). E o diâmetro do grupo (2,5) é 6.1.

Portanto, foram encontrados dois grupos que são $c_1 = \{amostra\ 2, amostra\ 5\}$ e $c_2 = \{amostra\ 1, amostra\ 3, amostra\ 4\}$ e respetivamente com os seus níveis 7,2 e 6,1. Agora o processo continuaria sobre o *cluster* com maior diâmetro (1,3,4) para encontrar o grupo dissidente prosseguindo o algoritmo até encontrar 5 clusters com 1 elemento cada. No final resultaria:

1º agrupamento: *cluster* (1,2,3,4,5) com diâmetro 8,7

2º agrupamento: *clusters* (2,5), (1,3,4). *Cluster* seleccionado (1,3,4) com diâmetro 7,2

3º agrupamento: *clusters* (2,5), (1), (3,4). *Cluster* seleccionado (2,5) com diâmetro 6,1

4º agrupamento: *clusters* (2),(5), (1), (3,4). *Cluster* seleccionado (3,4) com diâmetro 5,1

5º agrupamento: *clusters* (2),(5), (1), (3), (4). Fim do algoritmo.

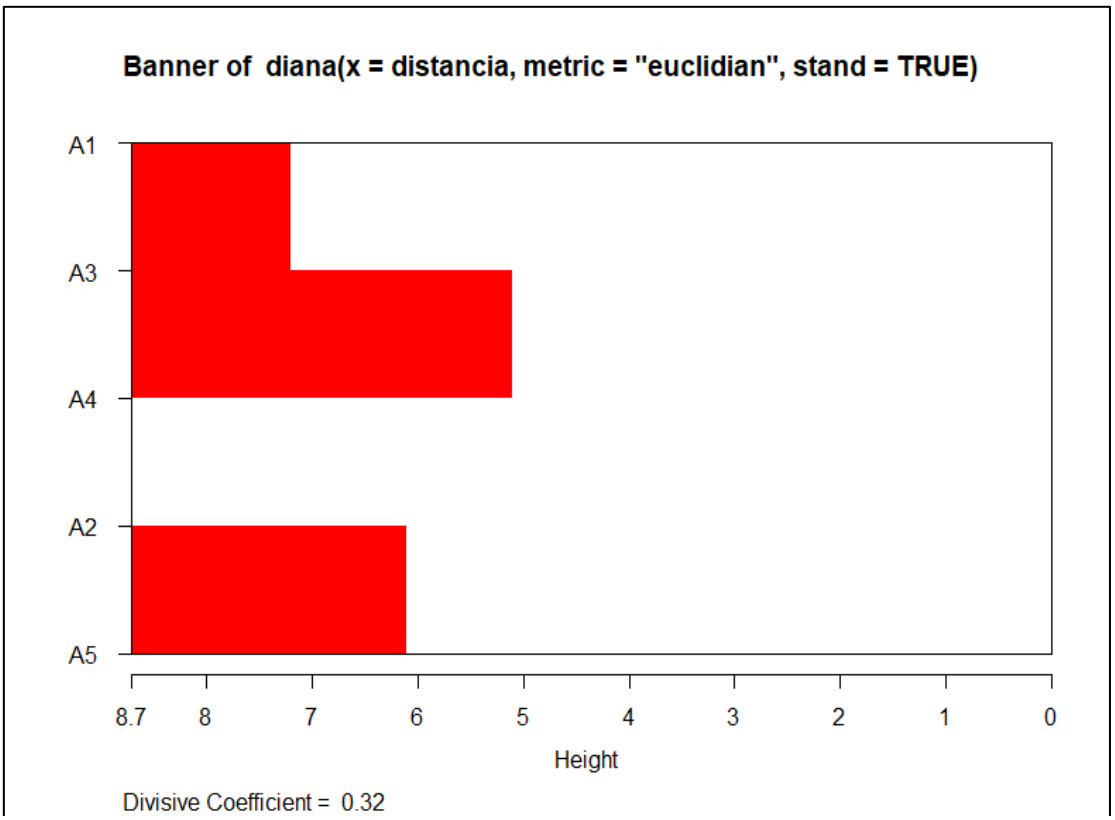

Figura 8:Exemplo do método divisivo Diana para o Exemplo 11

Na Figura 8 podemos calcular o valor do coeficiente divisivo, com respeito à linha de bandeira (*banner*), usando a equação da reta: $(y - y_0) = m(x - x_0)$, onde $(y_0, x_0) = (0,0)$ e $m = 1/8{,}7$, então $y = \frac{x}{8{,}7}$. Com base nos diâmetros calculados, observamos que:

- ✓ $y = 0{,}586$; $x = 5{,}1$ denotado por d_1
- ✓ $y = 0{,}70$; $x = 6{,}1$ denotado por d_2
- ✓ $y = 0{,}827$; $x = 7{,}2$ denotado por d_3

Logo, $CD = 1 - \frac{1}{5}(2 \times 0{,}586 + 2 \times 0{,}70 + 0{,}827) = 0{,}32$

2.3. Comparação de métodos aglomerativos

O objetivo desta secção é descrever um procedimento para comparar técnicas de aglomeração, que permita estabelecer, segundo algum critério, qual a técnica que apresenta

melhor desempenho entre os métodos considerados. Existem alguns critérios para avaliar o desempenho de técnicas de agrupamento. Uma das formas é baseada no coeficiente de correlação cofenética (definido por Sokal e Rohlf, 1962, citado em Reis, 2001) o qual é muito utilizado por taxonomistas numéricos. O uso deste coeficiente também é sugerido em Saraçli *et al* (2013).

O coeficiente de correlação cofenética mede o grau de ajuste entre a matriz de similaridade original e a matriz resultante da simplificação proporcionada pelo método de agrupamento e é determinado usando a fórmula do coeficiente de correlação linear de Pearson entre os elementos da matriz de distâncias originais e os correspondentes elementos da matriz de correlação cofenética, de acordo com a expressão:

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j>i}^{n}\left(c_{ij}-\bar{c}\right)\left(d_{ij}-\bar{f}\right)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j>i}^{n}\left(c_{ij}-\bar{c}\right)^2 \sum_{i=1}^{n-1}\sum_{j>i}^{n}\left(d_{ij}-\bar{f}\right)^2}}$$

em que;

$$\bar{c} = \frac{\sum_{i=1}^{n} c_i}{n}, \text{ e } \quad \bar{f} = \frac{\sum_{i=1}^{n} d_i}{n}$$

c_{ij} é o valor da distância entre os indivíduos i e j na matriz cofenética; d_{ij} é o valor de distância entre os mesmos indivíduos na matriz original de distâncias que podemos chamar fenética e n é a dimensão da matriz. A matriz resultante de qualquer método aglomerativo chamamos de matriz cofenética. O valor do coeficiente varia entre -1 e +1, com o valor zero a significar que não existe correlação entre os indivíduos. Para encontrar bons resultados da correlação cofenética, vários autores sugerem diferentes valores limiares; alguns indicam valores próximos de 0,7 e outros próximos de 0,8. Não há, portanto, consenso, sendo esse valor limiar subjetivo. Assim, neste trabalho, consideramos que um método produz um bom desempenho se o valor do coeficiente de correlação cofenética é maior ou igual 0,7.

Exemplo 12: Considere-se a matriz de distância abaixo

$$f = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & (5{,}1) & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Esta matriz corresponde à matriz de distâncias original f. Pretende-se avaliar o desempenho da aplicação do método de vizinho mais próximo. Este método conduz aos seguintes resultados agrupados:

Passo	**Distâncias e Grupos**
1	$d_{(43)} = 5{,}1$
2	$d_{(34)5} = \min\left(d_{(3,5)}\,;\, d_{(4,5)}\right) = 5{,}7$
3	$d_{(435)2} = \min\left(d_{(3,2)}, d_{(4,2)} \text{ e } d_{(5,2)}\right) = 6{,}1$
4	$d_{(4352)1} = \min\left(d_{(1,2)}, d_{(1,3)}, d_{(1,4)}, d_{(1,5)}\right) = 6{,}6$

Assim, obtemos a matriz cofenética *c* composta por:

$$c = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 6{,}6 & 0 & & & \\ 6{,}6 & 6{,}1 & 0 & & \\ 6{,}6 & 6{,}1 & 5{,}1 & 0 & \\ 6{,}6 & 6{,}1 & 5{,}7 & 5{,}7 & 0 \end{bmatrix} \end{array}$$

Assim, tem-se a correspondência entre elementos homólogos das matrizes *c* e *f* como se indica na tabela seguinte:

F	C
7,3	6,6
7,2	6,6
6,6	6,6
8,7	6,6
6,9	6,1
8,4	6,1
6,1	6,1
5,1	5,1
5,7	5,7
6,3	5,7

Para obter o coeficiente de correlação cofenética, deve-se calcular os valores da média e desvio padrão das matrizes fenética e cofenética.

- ✓ A média dos elementos da matriz fenética:

$$\bar{f} = \frac{\sum_{i=1}^{n} f_i}{n} = \frac{7,3 + 7,2 + 6,6 + 8,7 + 6,9 + 8,4 + 6,1 + 5,1 + 5,7 + 6,3}{10} = 6,75$$

- ✓ O desvio padrão dos elementos da matriz fenética:

$$S_F = \sqrt{\frac{\sum_{i=1}^{n} \left(f_i - \bar{f}\right)^2}{n-1}} = 1,14$$

- ✓ A média dos elementos da matriz cofenética

$$\bar{c} = \frac{\sum_{i=1}^{n} c_i}{n} = \frac{6,6 + 6,6 + 6,6 + 6,6 + 6,1 + 6,1 + 6,1 + 5,1 + 5,7 + 5,7}{10} = 6,1$$

- ✓ O desvio padrão dos elementos da matriz cofenética:

$$S_C = \sqrt{\frac{\sum_{i=1}^{n}(c_i - \bar{c})^2}{n-1}} = 0.53$$

A covariância entre as duas variáveis (variável fonética e variável cofenética) é definida por:

$$Cov_{(F,C)} = \frac{1}{n-1}\left(\sum_{i=1}^{n} f_i c_i - \frac{\sum_{i=1}^{n} f_i \sum_{i=1}^{n} c_i}{n}\right)$$

$$Cov_{(F,C)} = \frac{1}{10-1}\left(421{,}63 - \frac{68{,}3 \times 61{,}2}{10}\right) = 0{,}423$$

Por isso com base nesse resultado, podemos calcular o coeficiente da correlação cofenética dado por:

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j>i}^{n}(c_{ij} - \bar{c})(f_{ij} - \bar{f})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j>i}^{n}(c_{ij} - \bar{c})^2 \quad \sum_{i=1}^{n-1}\sum_{j>i}^{n}(f_{ij} - \bar{f})^2}} = \frac{Cov(F,C)}{S_F \times S_C} = \frac{0{,}423}{1{,}14 \times 0{,}53} = 0{,}705$$

Logo, $r_{cof} = 0{,}705 \geq 0{,}7$, podemos concluir que o método do vizinho mais próximo sobre a matriz de distâncias dada foi adequado para resumir a informação do conjunto de dados ou seja, esse método apresenta um bom desempenho.

2.4. Métodos não-hierárquicos

Um método não-hierárquicos é uma técnica de agrupamento onde os dados são divididos em *k* partições ou grupos, sendo que cada partição representa um *cluster*. Os números de *clusters* deve ser conhecido a priori. Os métodos não-hierárquicos têm a vantagem de poderem ser aplicados a conjuntos de dados de elevada dimensionalidade ou cardinalidade sem afetar grandemente a sua eficiência computacional, ao contrário das técnicas hierárquicas.

Nos métodos não hierárquicos começa-se a partir de: uma partição inicial de itens em grupos ou um conjunto inicial de pontos de sementes, que formarão os núcleos dos *clusters* que definem a partição inicial. As melhores escolhas para definir esta partição inicial esperam-se que não sejam enviesadas. Uma maneira de começar é selecionar aleatoriamente os pontos de semente entre os itens ou dividir aleatoriamente os itens em grupos iniciais.

Iremos considerar dois métodos não-hierárquicos: o método k-*means* (ou k-médias) e método k-medóides

Metodo k-means

O termo *k-means* é um metodo de agrupamento que foi introduzindo por MacQueen (1967), como é citado por Johnson et al. (2014); este sugere atribuir a cada um de seus itens para o *cluster* com o centróide (média) mais próximo. O algoritmo k-médias é um método de partição, cuja finalidade é minimizar a soma dos quadrados das distâncias ao centro do grupo respetivo, ou seja, visa a partição de n observações em k *clusters*.

O algoritmo de agrupamento *k-means* parte de um número k de *clusters* e é dada por:

1. Escolher o número k de itens de forma aleatória e declará-los como centróides iniciais.
2. Determinar o centróide mais próximo de cada um dos pontos e, em seguida, atribuir o ponto para o agrupamento associado com o centróide posicionado a uma distância menor do ponto.
3. Atualizar o centróide de cada *cluster* com base dos itens presentes nesse *cluster*. Geralmente, o novo centróide atualizados será a média de todos os pontos do *cluster* .
4. Repita o passo 2 e 3 até que não ocorram mais reatribuições.

Uma solução frequentemente utilizada para identificar o número ótimo de *cluster* é usando a chamada *regra do cotovelo* e envolve a observação de um conjunto de possíveis números de *clusters* em relação à forma como minimizam a soma de quadrados dentro do *cluster* (Kodinariya et al, 2013). Por outras palavras, a regra do cotovelo examina a

dissimilaridade dentro do *clusters* em função do número de *clusters*. Uma vez que o valor do critério da solução (soma de quadrados dentro dos grupos) tenderá a diminuir com o aumento sucessivo no número de *clusters*, a regra do cotovelo indica que o número "óptimo" de *clusters* é identificado onde se observa um "pico" na linha do gráfico

Exemplo 13: Relativamente à matriz original do Exemplo 4, assumimos que $k = 2$. O objetivo é dividir esses itens em *clusters* de modo que os itens dentro de um *cluster* sejam mais próximos uns dos outros do que os itens em diferentes *clusters*. Para implementar o método *k-means*, partimos arbitrariamente os itens em dois *clusters*, por exemplo: $C_1 = \{3,4,5\}$ e $C_2 = \{1,2\}$ e calculamos as coordenadas do centróide (média) de cada *cluster*.

Passo 1: Calcular a média dos grupos:

	Média de cada variável por *cluster*					
Grupo	$\overline{V1}$	$\overline{V2}$	$\overline{V3}$	$\overline{V4}$	$\overline{V5}$	$\overline{V6}$
C_1	7	3,6	2,3	2,3	5,6	8
C_2	3	3	2	6	6	7,5

Passo 2: Calcular a distância euclidiana de cada item aos centróides de cada grupo e reatribuir cada item ao grupo mais próximo. Considere os *clusters* iniciais definidos pelas coordenadas dos centróides da tabela acima. Calculando a distância euclidiana de cada objeto ao centróide dos grupos, obtém-se:

$$d_{(1,(345))} = \sqrt{(1-7)^2 + (2-3{,}6)^2 + \cdots + (6-8)^2} = 6{,}80$$

$$d_{(1,(12))} = \sqrt{(1-3)^2 + (2-3)^2 + \cdots + (6-7{,}5)^2} = 3{,}64$$

Da mesma maneira para os restantes indivíduos, tem-se a seguinte tabela:

Distância em cada elemento ao centróide do grupo	Valor
$d_{(1,(345))}$	6,80
$d_{(1(,12))}$	3,64
$d_{(2,(345))}$	6,42
$d_{(2,(12))}$	3,64
$d_{(3,(345))}$	2,98
$d_{(3,(12))}$	6,02
$d_{(4,(345))}$	3,27
$d_{(4,(12))}$	6,65
$d_{(5,(345))}$	3,61
$d_{(5,(12))}$	6,57

Classificar os objetos com base na distância mínima

Objeto	{345} (1)	{12} (2)	Mínimo	*Cluster*
1	6,80	3,64	3,64	2
2	6,42	3,64	3,64	2
3	2,98	6,02	2,98	1
4	3,27	6,65	3,27	1
5	3,61	6,57	3,61	1

Na tabela acima verifica-se que não há objetos que sejam movidos, portanto o processo de agrupamento fica concluído. Os cinco objetos dados ficam particionados nos *clusters* $C_1 = \{3,4,5\}$ e $C_2 = \{1,2\}$.

Quando existe algum elemento que está mais próximo de outro *cluster*, então ele deve ser retirado do grupo onde que está e associado ao outro grupo que lhe está mais próximo. No fim recalculam-se os centróides dos novos *clusters* e as distancias de cada elemento os

centróides, e assim sucessivamente até não serem encontrados objetos que se movam entre grupos.

Método k-medóides

O método k-medóides habitualmente designado por *Partioning Around Medoid* (PAM) foi desenvolvido por Rousseeuw (1987) como referido em Kaufman *et al* (2005). O k-medóides minimiza a soma de dissimilaridades entre pontos rotulados como sendo de um *cluster* e um ponto designado como o centro desse *cluster* ; é diferente do k-médias que pretende minimizar a somas de quadrados dos erros.

O PAM também é uma técnica de divisão de *clusters* que agrupa o conjunto de dados de n objetos em k *clusters*, com k conhecido *a priori*. Uma ferramenta útil para determinar k é o índice de silhueta.

O algoritmo k- medóides funciona de modo semelhante ao k-médias, mas o cálculo do centróide de cada grupo é diferente. No k-medóides, é entre os objetos de cada grupo que é eleito o centróide, enquanto no k-médias deve ser calculado um centro para o grupo. Assim, o algoritmo de PAM é como segue:

1. Escolher aleatoriamente k elementos do conjunto de dados como elementos representativos iniciais.
2. Atribuir cada elemento ao *cluster* como elemento representativo mais perto.
3. Escolher aleatoriamente um elemento não considerado representativo.
4. Calcular o custo (s) do elemento representativo escolhido no passo 3.
5. Se $s < 0$ troca o elemento representativo escolhido no passo 3.
6. Repete os passos anteriores até não haver mudança.

Exemplo 14: Relativamente ao Exemplo 4, aplicaremos o método k-medóides com a distância euclidiana; A matriz distância é dada por:

$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{bmatrix}
0 & & & & \\
7{,}3 & 0 & & & \\
7{,}2 & 6{,}9 & 0 & & \\
6{,}6 & 8{,}4 & 5{,}1 & 0 & \\
8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0
\end{bmatrix}
\end{array}
$$

Assumido que $k = 2$, e escolhendo aleatoriamente os elementos 1 e 5 como os medóides iniciais, temos:

Indivíduo (i)	$\mathbf{d_{i,1}(1)}$	$\mathbf{d_{i,5}(2)}$	$\mathbf{min\ (d_{i,1}; d_{i,5})}$	**Medóide mais próximo**
1	0	8,7	0	1
2	7,3	6,1	6,1	2
3	7,2	5,7	5,7	2
4	6,6	6,3	6,3	2
5	8,7	0	0	2
			Média = 3,62	

Com base na tabela acima verificamos que os indivíduos $2, 3$ e 4 ficam agrupados ao medóide 5, pois estão mais próximos deste, e nenhum indivíduo fica agrupado ao medóide 1. A média das similaridades mínimas calculadas na tabela acima representa a qualidade dos grupos encontrados. Portanto, quanto menor esse valor, melhor a qualidade dos grupos. Essa média representa o custo s na mudança dos medóides.

Para verificar a necessidade da mudança dos medóides, seleciona-se aleatoriamente um outro indivíduo, por exemplo, o indivíduo 3, e calcula-se o custo de substituir o medóide 3 por 5.

Indivíduo (i)	**$d_{i,1}$ (1)**	**$d_{i,3}$ (2)**	**min ($d_{i,1}$; $d_{i,3}$)**	**Medóide mais próximo**
1	0	7,2	0	1
2	7,3	6,9	6,9	2
3	7,2	0	0	2
4	6,6	5,1	5,1	2
5	8,7	5,7	5,7	2
Média = 3,54				

Assim, verificamos que os indivíduos $2, 4$ e 5 são agrupados ao medóide 3, pois estão mais próximo desse medóide e nenhum elemento (indivíduo) agrupado ao medóide 1. Calculando o custo de mudança do medóide 3 pelo 5 temos:

$$s_{3,5} = s_{1,3} - s_{1,5} = 3{,}54 - 3{,}62 = -0{,}08$$

Como o custo é menor que zero, então o medóide 5 é substituído pelo medóide 3. O algoritmo prossegue selecionando novos não-medóide verificando a necessidade de substituir os medóides. Essa análise é feita para todos os pares de elementos.

Procedendo da mesma maneira para outros indivíduos, substituindo os não-medóides para medóides, tem-se em resumo os seguintes resultados:

Medóides	(1-2)	**(1-3)**	(1-4)	(1-5)	(2-3)	(2-4)	(2-5)	(3-4)	(3-5)	(4-5)
Média	3,92	**3,54**	3,74	3,62	3,6	3,56	3,86	3,84	3,68	3,56

Verifica-se que os medóides 1 e 3 possuem a menor média; portanto, esses são os medóides finais e serão utilizados para formar os grupos. Logo, os indivíduos $2, 4\ e\ 5$ são agrupados ao medóide 3. E nenhum indivíduo é agrupado ao medóide 1.

Critérios de formação de clusters para dados contínuos

Os critérios de agrupamento mais utilizados, na análise de uma matriz de dados contínuos $X_{n\times p}$, usam a composição de matriz de dispersão T definida por:

$$T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}\right) \times \left(x_{ij} - \bar{x}\right)^T$$

em que x_{ij} é o vetor de dimensão p das observações do objeto i no grupo j e $\bar{x}$ é o vetor de dimensão p das médias das p variáveis nos n objetos de cada variável; ou seja

$$\bar{x} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{k} n_j}$$

Esta matriz de dispersão total pode ser dividida por:

- Matriz de dispersão dentro do grupo (W), dada por:

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(\bar{x}_{ij} - \bar{x}\right) \times \left(\bar{x}_{ij} - \bar{x}\right)^T$$

  em que $\bar{x}_j$ é o vetor de dimensão p das médias variáveis dentro do grupo j.

- Matriz de dispersão entre grupos (B), dada por:

$$B = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right) \times \left(\bar{x}_j - \bar{x}\right)^T$$

Então

$$T = B + W$$

onde T, W e B são as matrizes associados à variabilidade total dos dados, à variabilidade dentro dos grupos e à variabilidade entre grupos, respetivamente.

Para os dados univariados ($p = 1$), a equação $T = B + W$ representa a divisão da soma de quadrados total de uma variável na soma de quadrados dentro e entre grupos, fundamental na análise de variância.

No caso univariado, um critério natural para um agrupamento seria escolher a partição correspondente ao valor mínimo da soma de quadrados dentro dos grupos ou, de forma equivalente, ao valor máximo da soma de quadrados entre grupos. Assim, quanto maior a homogeneidade dentro dos grupos, maior é a separação entre os grupos.

a. Minimização do traço (W)

No caso multivariado $(p > 1)$, generalizar o caso sugerido da análise univariada, apesar do critério , $T = B + W$ não é tão claro quando $p = 1$.

Para determinar as três somas de quadrados anterior, referidas relativamente às p variáveis, precisamos da soma dos elementos da diagonal principal destas matrizes. Estas somas quadradas são definidas por: trT, trW e trB.

Uma extensão óbvia, para o caso multivariado, é a minimização das somas de quadrados dentro dos grupos, sobre todas as variáveis. Isto é, para minimizar o traço (W) que é, naturalmente, equivalente a maximizar traço (B).

Minimizar o traço (W) é equivalente a minimizar a somas dos quadrados das distâncias euclidianas entre indivíduos e respetiva média de grupo, isto é:

$$E = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)\times\left(x_{ij} - \bar{x}_j\right)^T = \sum_{j=1}^{k}\sum_{i=1}^{n_j} d_{ij,j}^2 = \sum_{l=1}^{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ijl} - \bar{x}_{jl}\right)^2$$

onde $d_{ij,j}$ é a distância euclidiana entre o i indivíduo no grupo j e a média do grupo j. Um critério fundamental também pode ser derivado com base na matriz de distância:

$$E = \sum_{j=1}^{k}\frac{1}{2n_j}\sum_{i=1}^{n_j}\sum_{v=1,v\neq i}^{n_j} d_{ji,vj}^2$$

onde $d_{ij,j}$ é a distância euclidiana entre o indivíduo i e indivíduo v no grupo j e a média do grupo j. Assim, a minimização do traço (W) é equivalente à minimização do critério de falta de homogeneidade da distância euclidiana que é usada por Ward para o processo hierárquico de formação *cluster*.

b. Minimização de determinante (W)

Na análise de variância multivariada, um dos testes para a diferença de novos vetores média dos grupos é baseada nos quocientes dos determinantes das matrizes de variabilidade total e matriz de variabilidade dentro dos grupos, $\frac{\det(T)}{\det(W)}$. Grandes valores de $\frac{\det(T)}{\det(W)}$ indicam que os vetores médios não são idênticos dos grupos. Dado que para todas as partições dos

indivíduos em grupos, T permanece a mesma, a maximização de $\frac{\det(T)}{\det(W)}$ é equivalente à minimização de $\det(W)$. Este critério, como referido por (Everitt, 2011), foi estudado por Marriott (1971, 1982),

c. A maximização do traço (BW^{-1})

Esta função é um critério de teste adicional utilizado no contexto da análise de variância multivariada, que usa o traço (BW^{-1}). Grandes valores do traço (BW^{-1}) indicam que os vetores médios não são idênticos dos grupos. Com base neste critério, a melhor partição será a que corresponde à maximização do traço da matriz obtida a partir do produto da matriz de dispersão entre os grupos e o inverso da matriz de dispersão dentro dos grupos. Quanto maior é o traço (BW^{-1}), e quanto menor $|W|$, maior é a diferença entre as médias dos grupos.

2.5 Medidas de validação de *clusters*

As medidas de validação são processos para avaliar os resultados ou obter o número ótimos de *clusters*. Nas medidas de validação existem três tipos de critérios principais para validar os *clusters*, nomeadamente*:* Critério de validação externa, Critério de validação interna e Critério de validação *relativo* (Jain *et al*, 2008; Wunsch II *et al*, 2008). Mas, importa saber que, conforme refere Brock *et al* (2008), cada conjunto formado tem um conjunto característico de medidas para analisar a validade de um agrupamento. Nesta dissertação focam-se dois critérios de validação: validação interna e estabilidade.

2.5.1. Medida de validação interna

A validação interna utiliza informação (interna) dos dados para avaliar a qualidade dos *clusters* formados. A validação interna inclui as seguintes medidas: conectividade, silhueta e índice de Dunn. Entre elas, a conectividade deve ser minimizada, enquanto o índice de Dunn e índice de silhueta devem ser maximizados.

a) Conectividade

Este índice indica o grau de conectividade dos agrupamentos determinado pelos vizinhos mais próximos e deve ser minimizado.

Defina $nn_{i(j)}$ como o $j-ésimo$ vizinho mais próximo da observação i, e seja $x_{i,nn_{i,(j)}} = 0$, se i e $nn_{i(j)}$ estiverem no mesmo *cluster*, e $x_{i,nn_{i,(j)}} = 1 \,/\, j$ caso contrário. Então, para uma partição de agrupamento particular, $C = (C_1, \cdots, C_k)$, das n observações em k $clusters$ disjuntos, a conectividade é definida por:

$$Conn(C) = \sum_{i=1}^{n} \sum_{j=1}^{L} x_{i,nn_{i,(j)}}$$

onde L é um parâmetro que determina o número de vizinhos, que contribuem para a medida de conectividade. Nestas condições, a conectividade varia entre $[0, \infty]$.

Exemplo 15: Considere-se a matriz de distâncias seguinte:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & 5{,}1 & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Passo 1: Suponha-se que uma dada técnica de agrupamento conduziu aos seguintes 2 c*lusters:* $c_1 = \{1, 3, 4\},\ c_2 = \{2, 5\}$.

Passo 2: Vamos agora construir a matriz com a informação dos vizinhos mais próximos $nn_{i(j)}$, onde: $nn_{i(j)}$ é o $j-ésimo$ vizinho mais próximo da observação i ; $i = 1, 2, 3, 4, 5$ representa os números de observações e *j* representa a ordem (pode ser $1\underline{0}, 2\underline{0}, 3\underline{0}, etc.$).

Fixemo-nos na observação $i = 1$. Tomando os seus vizinhos mais próximos, 2, 3, 4 e 5 e as suas distâncias, identificamos a ordem *j* resultando:

Observação do j − ésimo	2	3	4	5
Distância $(1, j)$	7,3	7,2	6,6	8,7
Ordem j	$3^{\underline{0}}$	$2^{\underline{0}}$	$1^{\underline{0}}$	$4^{\underline{0}}$

Logo, $nn_{1(j)} = [4\ 3\ 2\ 5]$ é um vetor com a indicação dos números das observações na posição $j - ésimo$ mais perto de $i = 1, com\ j = 1,2,3,4$. Procedendo da mesma maneira para $i = 2, 3, 4\ e\ 5$, podemos construir a matriz:

$$nn_{i,(j)} = \begin{bmatrix} 4 & 3 & 2 & 5 \\ 5 & 3 & 1 & 4 \\ 4 & 5 & 2 & 1 \\ 3 & 5 & 2 & 1 \\ 3 & 2 & 4 & 1 \end{bmatrix}$$

Passo 3: Vamos construir a matriz dos valores $[x_i, nn_{i,(j)}]$. Quando o valor *i* e o valor $nn_{i(j)}$ pertencem ao mesmo c*luster*, colocar o valor zero e, caso contrário, colocar o valor $\frac{1}{j}$. Por exemplo, considerando os c*lusters* $c_1 = \{1, 3, 4\}$ e $c_2 = \{2, 5\}$, como 1 e 4 estão no mesmo cluster, o valor de $[x_i, nn_{i,(j)}]$ para o par (1,4) será zero; e como 1 e 5 não estão no mesmo cluster e a observação 5 está próximo de 1 na ordem $j = 4^{\underline{0}}$, então o par (1,5) se terá $\frac{1}{j} = \frac{1}{4}$. Procedendo desta forma para todos os pares possíveis, obtemos a matriz:

$$\left[x_i, nn_{i(j)}\right] = \begin{array}{c} \\ 1 \\ 3 \\ 4 \\ 2 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 3 & 4 & 2 & 5 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1/3 & 1/4 \\ 0 & 0 & 0 & 1/3 & 1/2 \\ 0 & 0 & 0 & 1/4 & 1/2 \\ 1/3 & 1/2 & 1/4 & 0 & 0 \\ 1/4 & 1 & 1/3 & 0 & 0 \end{bmatrix} \end{array}$$

Logo, a soma de todos os elementos desta matriz corresponde ao valor da conectividade, isto é:

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i,(j)}} = 4{,}82$$

b) *Índice de silhueta*

O índice da silhueta é a média da largura de silhueta de cada observação. O valor de silhueta mede o grau de confiança na atribuição de um *cluster* a uma observação particular, sendo que para observações bem agrupadas teremos valores próximos de 1 e observações não bem agrupadas teremos valores próximos de -1. Portanto, varia entre $[-1; 1]$, e deve ser maximizada. Para observação i, o valor da largura de silhueta é definido por:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$$

onde a_i é a distância média entre i e todas as outras observações no mesmo c*luster* e b_i é a distância média entre i e as observações no c*luster* vizinho mais próximo, i.e.

$$a_i = \frac{1}{n(C_i)} \sum_{j \epsilon C_i} d(i,j)$$

$$b_i = \sum_{j \epsilon C_{k,}} \min_{C_k \epsilon C \setminus C_i} \frac{1}{n(C_k)} d(i,j)$$

onde C_i é o agrupamento que contém a observação i, $dist(i,j)$ é a distância (Ex: euclidiana, Manhattan, etc), entre as observações i e j, e $n(C)$ é a cardinalidade do agrupamento C .

Observe-se que a largura de silhueta também pode ser escrita, de forma equivalente, na forma:

$$S(i) = \begin{cases} 1 - \frac{a_i}{b_i}; & se\ a_i < b_i \\ 0\ ; & se\ a_i = b_i \\ \frac{b_i}{a_i} - 1; & se\ a_i > b_i \end{cases}$$

A largura de silhueta é simplificada no agrupamento final presente na equação (2.5), e o índice de silhueta é obtido pela média das silhuetas dos objetos do conjunto de dados:

$$CS = \frac{1}{n} \sum_{i=1}^{n} S(i) \qquad (2.5)$$

onde n é os números de observações.

A Tabela 8 mostra valores de referência empiricamente aceites para interpretar o coeficiente de silhueta (CS) (Kaufam et al., 2005).

CS	Interpretação
0,71-1,00	Estrutura forte
0,51-0,70	Estrutura razoável
0,26-0,50	Estrutura fraca e pode ser artificial
< 0.25	Nenhuma estrutura substancial

Tabela 8: Interpretação do coeficiente de silhueta (CS)

Exemplo 16: Considere-se a matriz de distâncias seguinte:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & 5{,}1 & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Passo 1: Suponhamos que uma técnica de agrupamento conduziu a dois c*lusters*: $c_1 = \{1, 3, 4\}$ e $c_2 = \{2, 5\}$.

Passo 2: Vamos calcular as silhuetas em cada observação $i = 1,2,3,4,5$. Fixando a observação $i = 1$, temos:

$$a_1 = \frac{1}{n(C_1)} \sum_{j \epsilon C_1} dist(1, j) = \frac{1}{2}(d_{14} + d_{13}) = \frac{1}{2}(6{,}6 + 7{,}2) = 6{,}9$$

$$b_i = \min_{C_2 \epsilon C \setminus C_1} \sum_{j \epsilon C_2} \frac{1}{n(C_2)} dist(1, j) = \frac{1}{2}(d_{12} + d_{15}) = \frac{1}{2}(7{,}3 + 8{,}7) = 8$$

$$S(1) = \frac{b_1 - a_1}{max\ (b_1, a_1)} = \frac{8 - 6{,}9}{8} = 0{,}13$$

Da mesma maneira para $i = 2, 3, 4\ e\ 5$, no fim podemos resumir os resultados na tabela seguinte:

Observação i	a_i	b_i	$S(i)$
1	6,9	8	0,13
2	6,1	7,5	0,19
3	6,15	6,25	0,016
4	5,85	7,25	0,20
5	6,1	6,9	0,11

Passo 3: Vamos reunir os números das silhuetas e calcular as médias que pertencem a cada *cluster*

Cluster	**Silhueta**	**Média das silhuetas**
$\mathbf{C_1}$	$\{S_1, S_3, S_4\}$	0,12
$\mathbf{C_2}$	$\{S_2, S_5\}$	0,15
Média das médias das silhuetas		$\mathrm{CS} = \frac{3 \times 0{,}12 + 2 \times 0{,}15}{5} = 0{,}13$

Logo, o valor do Coeficiente de Silhueta é 0,13, o que significa que nenhuma estrutura substancial foi encontrada com o agrupamento considerado $C_1 = \{1, 3, 4\}$ e $C_2 = \{2, 5\}$.

c) *Índice de Dunn*

O índice de Dunn é a razão entre a menor distância entre as observações que não estão no mesmo *cluster* e a maior distância intra-*clusters*. A sua fórmula de cálculo é:

$$D(C) = \frac{\min\limits_{C_k, C_l \epsilon C; C_k \neq C_l} \left(\min\limits_{i \epsilon C_k, j \epsilon C_l} d(i,j) \right)}{\max\limits_{C_m \epsilon C} diam(C_m)}$$

onde $diam(C_m)$ é a distância máxima entre observações no conjunto C_m. O índice Dunn varia entre $[0, +\infty[$ e deve ser maximizado.

Exemplo 17: Considere-se a matriz de distâncias seguinte

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & 5{,}1 & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

Pelo método vizinho mais-afastado obtemos dois grandes grupos, nomeadamente, o grupo $c_1 = \{1, 3, 4\}$ e $c_2 = \{2, 5\}$. Vamos calcular o índice de Dunn.

Ora,

$$Min\ d(C) = \min(d_{21},\ d_{23}, d_{24}, d_{51}, d_{53}, d_{54}) = \min(7{,}3; 6{,}9; 8{,}4; 8{,}7; 5{,}7; 6{,}3) = 5{,}7 \text{ e}$$

$$Max\ diam(C_m) = \max(d_{25}\ , d_{13}, d_{14}) = \max(6{,}1; 7{,}2; 6{,}6) \ = 7{,}2$$

Logo,

$$D(C) = \frac{Min\ d(C)}{Max\ diam(C_m)} = \frac{5{,}7}{7{,}2} = 0{,}79$$

2.5.2. Medida de estabilidade dos *Clusters*

Este tipo de medidas tem como objetivo avaliar a estabilidade dos resultados dos *clusters*. Estas medidas funcionam bem se os dados são altamentes correlacionados. Iremos considerar quatro medidas de estabilidade: Proporção média de não sobreposição (*average proportion of non-overlap*, *APN)*, distância média (*average distance*, *AD)*, distância média entre centros de *clusters* **(***Average distance between means*, *ADM)* e figura de mérito (*figure of merit*, *FOM).* Em todas estas medidas, a média é tomada em todas as colunas excluídas, e todas as medidas devem ser minimizadas.

a. Proporção média de não sobreposição *(APN)*

A APN mede a proporção média de observações não colocadas no "mesmo" *cluster* em duas situações distintas.

Seja $C_{i,0}$ o *cluster* contendo a observação i resultante de um agrupamento original (com base em todos os dados disponíveis) e seja $C_{i,l}$ o *cluster* contendo a observação i onde o agrupamento é baseado no conjunto de todos os dados com a coluna l removida. A medida APN é definida por:

$$APN = \frac{1}{pn}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(1 - \frac{n(C_{i,l} \cap C_{i,0})}{n(C_{i,0})}\right)$$

onde $n(C)$ corresponde ao cardinal do cluster *C*. O APN varia entre o intervalo [0; 1], com valores próximos de zero correspondentes a resultados de agrupamento altamente consistentes.

Exemplo 18: Relativamente à matriz de dados do Exemplo 4, tem-se a seguinte matriz de distâncias:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 7{,}3 & 0 & & & \\ 7{,}2 & 6{,}9 & 0 & & \\ 6{,}6 & 8{,}4 & 5{,}1 & 0 & \\ 8{,}7 & 6{,}1 & 5{,}7 & 6{,}3 & 0 \end{bmatrix} \end{array}$$

- ✓ Pelo método aglomerativo do vizinho mais afastado, obtemos dois grandes grupos, $c_1 = \{1, 3, 4\}$ e $c_2 = \{2, 5\}$. Assim, $c_{1,0} = c_1$; $c_{2,0} = c_2$; $c_{3,0} = c_1$; $c_{4,0} = c_1$; $c_{5,0} = c_2$
- ✓ Eliminando a variável V1, e calculando a medida de distância e aplicando novamente o método hierárquico do vizinho mais afastado com a variável removida, obtemos dois grandes grupos, nomeadamente: $c_1 = \{2\}$ e $c_2 = \{1{,}3{,}4{,}5\}$; $c_{1,1} = c_2$; $c_{2,1} = c_1$; $c_{3,1} = c_2$; $c_{4,1} = c_1$; $c_{5,1} = c_1$
- ✓ Procede-se da mesma maneira removendo cada uma das restantes variáveis até V6. Os resultados obtidos estão resumidos na tabela seguinte:

No.	Variável V*l* removida	C*lusters* resultantes do método do vizinho mais afastado	*Cluster* de cada observação *i* $c_{i,l}$
1.	---	$c_1 = \{1, 3, 4\}$ e $c_2 = \{2, 5\}$.	$c_{1,0} = c_1$; $c_{2,0} = c_2$; $c_{3,0} = c_1$; $c_{4,0} = c_1$; $c_{5,0} = c_2$
2.	V1	$c_1 = \{2\}$ e $c_2 = \{1,3,4,5\}$	$c_{1,1} = c_2$; $c_{2,1} = c_1$; $c_{3,1} = c_2$; $c_{4,1} = c_1$; $c_{5,1} = c_1$
3.	V2	$c_1 = \{3,4,5\}$ e $c_2 = \{1,2\}$	$c_{1,2} = c_2$; $c_{2,2} = c_2$; $c_{3,1} = c_1$; $c_{4,2} = c_1$; $c_{5,2} = c_1$
4.	V3	$c_1 = \{2,5\}$ e $c_2 = \{1,3,4\}$	$c_{1,3} = c_2$; $c_{2,3} = c_1$; $c_{3,3} = c_2$; $c_{4,3} = c_2$; $c_{5,3} = c_1$
5.	V4	$c_1 = \{1\}$ e $c_2 = \{2,3,4,5\}$	$c_{1,4} = c_1$; $c_{2,2} = c_2$; $c_{3,4} = c_2$; $c_{4,4} = c_2$; $c_{5,4} = c_2$
6.	V5	$c_1 = \{3,4,5\}$ e $c_2 = \{1,2\}$	$c_{1,5} = c_2$; $c_{2,5} = c_2$; $c_{3,5} = c_1$; $c_{4,5} = c_1$; $c_{5,5} = c_1$
7.	V6	$c_1 = \{2,5\}$ e $c_2 = \{1,3,4\}$	$c_{1,6} = c_2$; $c_{2,6} = c_1$; $c_{3,6} = c_2$; $c_{4,6} = c_2$; $c_{5,6} = c_1$

✓ Finalmente, calculamos a proporção média de não sobreposição *(APN)* dada por:

$$APN = \frac{1}{30}\left\{\left(1 - \frac{n(C_{1,1} \cap C_{1,0})}{n(C_{1,0})}\right) + \left(1 - \frac{n(C_{1,2} \cap C_{2,0})}{n(C_{2,0})}\right) + \cdots + \left(1 - \frac{n(C_{5,6} \cap C_{5,0})}{n(C_{5,0})}\right)\right\}$$

$$\begin{aligned} APN = \frac{1}{30}\Big\{&\left(1 - \frac{3}{3}\right) + \left(1 - \frac{0}{3}\right) + \left(1 - \frac{3}{3}\right) + \left(1 - \frac{1}{3}\right) + \left(1 - \frac{1}{3}\right) + \left(1 - \frac{3}{3}\right) + \left(1 - \frac{1}{2}\right) \\ &+ \left(1 - \frac{1}{2}\right) + \left(1 - \frac{2}{2}\right) + \left(1 - \frac{2}{2}\right) + \left(1 - \frac{1}{2}\right) + \left(1 - \frac{2}{2}\right) + \left(1 - \frac{3}{3}\right) + \left(1 - \frac{2}{3}\right) \\ &+ \left(1 - \frac{3}{3}\right) + \left(1 - \frac{2}{3}\right) + \left(1 - \frac{2}{3}\right) + \left(1 - \frac{3}{3}\right) + \left(1 - \frac{3}{3}\right) + \left(1 - \frac{2}{3}\right) + \left(1 - \frac{3}{3}\right) \\ &+ \left(1 - \frac{2}{3}\right) + \left(1 - \frac{2}{3}\right) + \left(1 - \frac{3}{3}\right) + \left(1 - \frac{1}{2}\right) + \left(1 - \frac{1}{2}\right) + \left(1 - \frac{2}{2}\right) + \left(1 - \frac{2}{2}\right) \\ &+ \left(1 - \frac{1}{2}\right) + \left(1 - \frac{2}{2}\right)\Big\} = 0{,}238 \end{aligned}$$

b. Distância média (AD)

A medida *AD* mede a distância média entre as observações colocadas no mesmo c*luster*, com base nos dados completos e com base no agrupamento de dados com uma única coluna removida. Portanto, *AD* é definida por:

$$AD = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{1}{n(C_{i,0})n(C_{i,l})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right]$$

A *AD* varia entre $[0,\infty]$. Valores menores de *AD* são preferíveis.

Exemplo 19: Relativamente à tabela resumido do Exemplo 18 podemos calcular:

- $\frac{1}{n(C_{1,0})n(C_{1,1})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = \frac{1}{12}(d_{13} + d_{14} + d_{15} + d_{31} + d_{34} + d_{35} + d_{41} +$ *d43+d45=4,875*
- $\frac{1}{n(C_{1,0})n(C_{1,2})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = \frac{1}{6}(d_{12} + d_{31} + d_{32} + d_{41} + d_{42}) = 6{,}066$

Com a mesma maneira para outra medida, finalmente obtemos o valor:

Passo	$\frac{1}{n(C_{i,0})n(C_{i,l})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right]$	**Passo**	$\frac{1}{n(C_{i,0})n(C_{i,l})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right]$
1	$\frac{1}{n(C_{1,0})n(C_{1,1})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$	2	$\frac{1}{n(C_{2,0})n(C_{2,1})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 3{,}1$
	$\frac{1}{n(C_{1,0})n(C_{1,2})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 6{,}1$		$\frac{1}{n(C_{2,0})n(C_{2,2})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 5{,}5$
	$\frac{1}{n(C_{1,0})n(C_{1,3})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}2$		$\frac{1}{n(C_{2,0})n(C_{2,3})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 3{,}1$
	$\frac{1}{n(C_{1,0})n(C_{1,4})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}6$		$\frac{1}{n(C_{2,0})n(C_{2,4})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$
	$\frac{1}{n(C_{1,0})n(C_{1,5})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 6{,}1$		$\frac{1}{n(C_{2,0})n(C_{2,5})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 5{,}5$
	$\frac{1}{n(C_{1,0})n(C_{1,6})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}2$		$\frac{1}{n(C_{2,0})n(C_{2,6})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 3{,}1$
3	$\frac{1}{n(C_{3,0})n(C_{3,1})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$	4	$\frac{1}{n(C_{4,0})n(C_{4,1})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$
	$\frac{1}{n(C_{3,0})n(C_{3,2})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$		$\frac{1}{n(C_{4,0})n(C_{4,2})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$
	$\frac{1}{n(C_{3,0})n(C_{3,3})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}2$		$\frac{1}{n(C_{4,0})n(C_{4,3})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}2$
	$\frac{1}{n(C_{3,0})n(C_{3,4})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 5{,}6$		$\frac{1}{n(C_{4,0})n(C_{4,4})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 5{,}6$

	$\frac{1}{n(C_{3,0})n(C_{3,5})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$	$\frac{1}{n(C_{4,0})n(C_{4,5})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$
	$\frac{1}{n(C_{3,0})n(C_{3,6})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}2$	$\frac{1}{n(C_{4,0})n(C_{4,6})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}2$
5	$\frac{1}{n(C_{5,0})n(C_{5,1})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 6{,}2$	
	$\frac{1}{n(C_{5,0})n(C_{5,2})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 5{,}6$	
	$\frac{1}{n(C_{5,0})n(C_{5,3})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 3{,}1$	
	$\frac{1}{n(C_{5,0})n(C_{5,4})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 4{,}9$	
	$\frac{1}{n(C_{5,0})n(C_{5,5})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 5{,}6$	
	$\frac{1}{n(C_{5,0})n(C_{5,6})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = 3{,}1$	

Deste modo, resulta:

$$AD = \frac{1}{30}\sum_{i=1}^{5}\sum_{j=1}^{6}\frac{1}{n(C_{i,0})n(C_{i,l})}\left[\sum_{i\epsilon C_{i,0}, j\epsilon C_{i,l}} d(i,j)\right] = \frac{141{,}2}{30} = 4{,}706$$

c. Distância média entre médias (ADM)

A medida ADM calcula a distância média entre centros de agrupamento para observações colocadas no mesmo c*luster* num agrupamento obtido com base nos dados completos e num agrupamento obtido com base nos dados com uma única coluna removida. A *ADM* é definida por:

$$ADM = \frac{1}{np}\sum_{i=1}^{n}\sum_{l=1}^{p} d\left(\bar{x}_{C_{i,0}}, \bar{x}_{C_{i,l}}\right)$$

onde $\bar{x}_{C_{i,0}}$ é a média das observações no c*luster* que contêm a observação i, quando o agrupamento se baseia nos dados completos, e $\bar{x}_{C_{i,l}}$ é definido de forma semelhante sobre um agrupamento baseado em dados incompletos resultantes da remoção da coluna *l* da matriz de dados. *ADM* varia no intervalo $[0, +\infty[$. Valores menores para *ADM* são preferíveis.

Exemplo 20: Relativamente à tabela resumida no Exemplo 18, resultante da matriz de dados original do Exemplo 4, podemos calcular o valor de *ADM*.

Passo 1: Calculamos os centróides $\bar{x}_{C_{i,l}}$ e $\bar{x}_{C_{i,l}}$, para $i = 1$ e $l = 1$, e o valor de distância entre eles, $dist\left(\bar{x}_{C_{i,l'}}, \bar{x}_{C_{i,0}}\right)$. Concretamente, temos:

- $\bar{x}_{C_{1,0}} = \frac{1}{3}[(1;2;3;4;5;6) + (6;5;4;2;7;9) + (6;4;2;1;3;7)] =$
  $(4{,}3;\ 3{,}6;\ \ 3{,}0;\ \ 2{,}3;\ \ 5{,}0;\ \ 7{,}3)$
- $\bar{x}_{C_{1,1}} = \frac{1}{4}[(1;2;3;4;5;6) + (6;5;4;2;7;9) + (6;4;2;1;3;7) + (9;2;1;4;7;8)]$
  $= (5{,}5;\ 3{,}25; 2{,}5;\ 2{,}75;\ \ 5{,}5; 7{,}5)$

$$d\left(\bar{x}_{i,0}, \bar{x}_{C_{i,l}}\right) = \sqrt{(4{,}3 - 5{,}5)^2 + (3{,}6 - 3{,}25)^2 + \cdots + (7{,}3 - 7{,}5)^2} = 1{,}50$$

Passo 2: Procedendo da mesma maneira para os restantes, no fim resultam os valores de distância sumariados na seguinte tabela:

Passo	Medida de distância $d(\bar{x}_{C_{i,0}}, \bar{x}_{C_{i,l}})$	Passo	Medida de distância $d(\bar{x}_{C_{i,0}}, \bar{x}_{C_{i,l}})$
1	$d(\bar{x}_{C_{1,0}}, \bar{x}_{C_{1,1}}) = 1{,}50$ $d(\bar{x}_{C_{1,0}}, \bar{x}_{C_{1,2}}) = 4{,}21$ $d(\bar{x}_{C_{1,0}}, \bar{x}_{1,3}) = 0$ $d(\bar{x}_{C_{1,0}}, \bar{x}_{C_{1,4}}) = 4{,}24$ $d(\bar{x}_{C_{1,0}}, \bar{x}_{C_{1,5}}) = 4{,}21$ $d(\bar{x}_{C_{1,0}}, \bar{x}_{C_{1,6}}) = 1{,}50$	2	$d(\bar{x}_{C_{2,0}}, \bar{x}_{C_{2,1}}) = 3{,}04$ $d(\bar{x}_{C_{2,0}}, \bar{x}_{C_{2,2}}) = 4{,}35$ $d(\bar{x}_{C_{2,0}}, \bar{x}_{C_{2,3}}) = 0$ $d(\bar{x}_{C_{2,0}}, \bar{x}_{C_{2,4}}) = 3{,}22$ $d(\bar{x}_{C_{2,0}}, \bar{x}_{C_{2,5}}) = 4{,}35$ $d(\bar{x}_{C_{2,0}}, \bar{x}_{C_{2,6}}) = 0$
3	$d(\bar{x}_{C_{3,0}}, \bar{x}_{C_{3,1}}) = 1{,}50$ $d(\bar{x}_{C_{3,0}}, \bar{x}_{C_{3,2}}) = 2{,}93$ $d(\bar{x}_{C_{3,0}}, \bar{x}_{C_{3,3}}) = 0$ $d(\bar{x}_{C_{3,0}}, \bar{x}_{C_{3,4}}) = 4{,}52$ $d(\bar{x}_{C_{3,0}}, \bar{x}_{C_{3,5}}) = 2{,}93$ $d(\bar{x}_{C_{3,0}}, \bar{x}_{C_{3,6}}) = 0$	4	$d(\bar{x}_{C_{4,0}}, \bar{x}_{C_{4,1}}) = 1{,}50$ $d(\bar{x}_{C_{4,0}}, \bar{x}_{C_{4,2}}) = 2{,}93$ $d(\bar{x}_{C_{4,0}}, \bar{x}_{C_{4,3}}) = 0$ $d(\bar{x}_{C_{4,0}}, \bar{x}_{4,4}) = 4{,}52$ $d(\bar{x}_{C_{4,0}}, \bar{x}_{C_{4,5}}) = 2{,}93$ $d(\bar{x}_{C_{4,0}}, \bar{x}_{C_{4,6}}) = 0$
5	$d(\bar{x}_{C_{5,0}}, \bar{x}_{C_{5,1}}) = 4{,}28$ $d(\bar{x}_{C_{5,0}}, \bar{x}_{C_{5,2}}) = 4{,}99$ $d(\bar{x}_{C_{5,0}}, \bar{x}_{C_{5,3}}) = 0$ $d(\bar{x}_{C_{5,0}}, \bar{x}_{C_{5,4}}) = 3{,}22$ $d(\bar{x}_{C_{5,0}}, \bar{x}_{C_{5,6}}) = 0$		

Logo,

$$ADM = \frac{1}{30}\sum_{i=1}^{5}\sum_{l=1}^{6} d(\bar{x}_{C_{i,0}}, \bar{x}_{C_{i,l}}) = \frac{66{,}87}{30} = 2{,}229$$

d. Figura de mérito (FOM)

A medida *FOM* mede a variância intra-cluster média das observações na coluna excluída, onde o agrupamento é baseado nas amostras restantes (não eliminadas). Assim, a *FOM* calcula o erro médio com base nas médias de c*luster*. Para uma coluna de exclusão "l", a FOM é:

$$FOM(l, C) = \sqrt{\frac{1}{Nn}\sum_{k=1}^{K}\sum_{i\epsilon C_{k,(l)}} d(x_{i,l}, \bar{x}_{C_{k,(l)}})}$$

onde $x_{i,l}$ é o valor da observação na linha i, coluna l da matriz original de dados, e $\bar{x}_{C_{k,(l)}}$ é a média das observações que estão na coluna l e pertencente ao grupo C_k. Atualmente, a única distância para a *FOM* é a distância euclidiana. A FOM final é obtida multiplicando por um fator de ajustamento $\sqrt{\frac{n}{n-K}}$ para aliviar a tendência para diminuir à medida que o número de c*lusters K* aumenta. A pontuação *FOM* final é a média tomando as colunas removidas sendo dada por :

$$FOM_{final} = \left(\frac{1}{p}\sum_{l=1}^{p} FOM(l, C)\right)\sqrt{\frac{n}{n-K}}$$

Esta medida varia no intervalo $[0, +\infty[$, com valores menores indicando melhor desempenho.

Exemplo 21: Relativamente à tabela resumo do Exemplo 16, e relativa à matriz de dados original do Exemplo 4, podemos calcular o valor de FOM. Temos:

Passo 1: calcular o valor de $FOM(l, C)$ para $l = 1$ e para os dois c*lusters* $c_1 = \{2\}$ e $c_2 = \{1,3,4,5\}$:

- $c_1 = \{2\}$, tem-se:
  - ✓ $x_{2,1} = 5$ e $\bar{x}_{c_{1,1}} = 5$; logo, $d\left(x_{2,1}, \bar{x}_{C_{1,(1)}}\right) = \sqrt{(5-5)^2} = 0$
- $c_2 = \{1,3,4,5\}$, tem-se:
  - ✓ $x_{1,1} = 1$ e $\bar{x}_{c_{2,1}} = \frac{1+6+6+9}{4} = 5{,}5$; logo, $d\left(x_{1,1}, \bar{x}_{C_{2,(1)}}\right) = \sqrt{(1-5{,}5)^2} = 4{,}5$
  - ✓ $x_{3,1} = 6$ e $\bar{x}_{c_{2,1}} = \frac{1+6+6+9}{4} = 5{,}5$; logo, $d\left(x_{3,1}, \bar{x}_{C_{2,(1)}}\right) = \sqrt{(6-5{,}5)^2} = 0{,}5$

- ✓ $x_{4,1} = 6$ e $\bar{x}_{c_{2,1}} = \frac{1+6+6+9}{4} = 5{,}5$; logo, $d\left(x_{4,1}, \bar{x}_{C_{2,(1)}}\right) = \sqrt{(6-5{,}5)^2} = 0{,}5$
- ✓ $x_{5,1} = 9$ e $\bar{x}_{c_{2,1}} = \frac{1+6+6+9}{4} = 5{,}5$; logo, $d\left(x_{5,1}, \bar{x}_{C_{2,(1)}}\right) = \sqrt{(9-5{,}5)^2} = 3{,}5$

Logo, $FOM(1, C) = \sqrt{\frac{1}{5}\left[\sum_{i \epsilon C_{1,(1)}} d\left(x_{i,1}, \bar{x}_{C_{1,(1)}}\right) + \sum_{i \epsilon C_{2,(1)}} d(x_{i,1}, \bar{x}_{C_{2,(1)}})\right]} = 1{,}34$

Passo 2: Procede-se da mesma maneira para as outras colunas a remover, resultando a tabela seguinte:

Passos	**Coluna removida (l)**	***Clusters***	$\boldsymbol{FOM(l, C)}$
1.	$l = 1$	$c_1 = \{2\}$ e $c_2 = \{1,3,4,5\}$	1,34
2.	$l = 2$	$c_1 = \{3,4,5\}$ e $c_2 = \{1,2\}$	3,72
3.	$l = 3$	$c_1 = \{2,5\}$ e $c_2 = \{1,3,4\}$	1,89
4.	$l = 4$	$c_1 = \{1\}$ e $c_2 = \{2,3,4,5\}$	1,34
5.	$l = 5$	$c_1 = \{3,4,5\}$ e $c_2 = \{1,2\}$	1,21
6.	$l = 6$	$c_1 = \{2,5\}$ e $c_2 = \{1,3,4\}$	0,92
$\boldsymbol{FOM_{total}}$	-	-	10,42

Logo, o valor final é: $FOM_{final} = \left(\frac{1}{6}\sum_{l=1}^{6} FOM(l, C)\right)\sqrt{\frac{5}{5-2}} = \frac{10{,}42}{6}\sqrt{\frac{5}{5-2}} = 2{,}24$

Capítulo 3

Aplicações e análise de resultados

Neste capítulo é apresentado o tratamento estatístico efetuado, com aplicação das técnicas referidas anteriormente, a um conjunto de dados reais. O *software* utilizado no tratamento de dados foi o R, versão 3.3.2 (R Core Team, 2016).

3.1. Descrição dos dados

Neste capítulo é realizada uma Análise de *Clusters* aplicada a um conjunto de dados fornecidos pela Profª Doutora Cristina Gomes, do Departamento de Ciências Sociais, Políticas e de Território da Universidade de Aveiro. Estes dados focam a movimentação de pessoas que mudaram de residência, entre 2005 e 2011, quantificando quantos novos residentes passaram a ter cada concelho de Portugal em 2011. A informação sobre os concelhos foi agregada e convertida a nível dos 18 distritos e das 2 regiões autónomas (de agora em diante designados, abusivamente, por "distritos" para simplificação de escrita), em vez dos 308 concelhos, com o objetivo de ser viável a visualização dos agrupamentos através de dendrogramas. Estes dados são compostos por 20 variáveis, sendo essas variáveis relativas a contagens sobre quatro características: idade, género, habilitação literária e situação profissional. As 20 variáveis (de contagem) estão identificadas na Tabela 9 e descrevem quatro características.

Dado que se pretende um agrupamento de distritos (indivíduos) utilizar-se-á uma medida de dissimilaridade. A medida a usar foi a distância euclidiana. Foram aplicadas diferentes técnicas de agrupamento, métodos hierárquicos aglomerativos, divisivos e métodos de partição, e avaliada a qualidade dos agrupamentos providenciados por cada técnica usando o pacote "**clValid**" do R . Para mais detalhe sobre os dados completos e *script* do R consultar **anexos A e B,** respetivamente.

Código	Descrição de Variáveis
I1	Idade de 0-14 anos
I2	Idade de 15-24 anos
I3	Idade de 25-39 anos
I4	Idade de 40-64 anos
I5	Idade mais de 65 anos
M	Sexo masculino
F	Sexo feminino
T1	Posição de trabalho com atividade económico desempregado
T2	Posição de trabalho com atividade económico empregado
T3	Posição de trabalho inactivo
H1	Bacharelato
H2	Doutoramento
H3	EB $1^{\underline{0}}$ ciclo
H4	EB $2^{\underline{0}}$ ciclo
H5	EB $3^{\underline{0}}$ ciclo
H6	Ensino pós-secundário
H7	Ensino secundário
H8	Licenciatura
H9	Mestrado
H10	Nenhum

Tabela 9: Identificação das variáveis relativas aos dados estudados

3.2. Análise e comparação de técnicas hierárquicas

A Análise de *Clusters* tem com objetivo criar grupos homogéneos, mas é necessário escolher o número adequado de *clusters*. A aplicação de métodos hierárquicos permite a apresentação de resultados sob a forma de dendrogramas e assim providenciar uma técnica visual para identificar o número de *clusters* e os *clusters*. Para identificar quais os *clusters* a tomar, começamos este estudo por comparar os resultados de diferentes técnicas hierárquicas usando o coeficiente de correlação cofenética. Os resultados obtidos encontram-se na Tabela

10 tomando a matriz de dados seccionada pelas quatro características atrás mencionadas e ainda tomando todas as variáveis numa única matriz de dados.

Técnica Aglomerativa	Género	Situação de trabalho	Habilitação literária	Idade	Dados gerais
Vizinho mais próximo	0,7667	0,7667	0,655	0,6265	0,6835
Vizinho mais afastado	0,7628	0,7934	0,7339	0,4513	0,7540
Média	**0,7898**	**0,7964**	**0,7467**	**0,7525**	**0,7640**
Centróide	0,7871	0,7952	0,7254	0,6827	0,7484
Ward	0,7589	0,7839	0,7088	0,4932	0,7432

Tabela 10: Resultado do coeficiente correlação cofenética dos métodos hierárquicos aplicados sobre os dados

Geralmente um valor de correlação cofenética maior que 0,7 indica que a matriz de correlação cofenética, gerada a partir do dendrograma, representa uma boa simplificação da matriz de distâncias ou fenética. Portanto, da Tabela 10, resulta que, com base neste critério, para as quatro características e os dados agregados, o método da média seria uma boa escolha pois indicia um bom desempenho para formar os grupos.

a. Análise de dados da característica *Idade* por distrito

Nos casos dos dados das idades, o método da média, através do dendrograma (Figura 9), sugere dois grupos de distritos em que a Madeira aparece isolada dos restantes distritos (Santarém, Leiria, Beja, Guarda, Aveiro, Bragança, Coimbra, Faro, Évora, Viseu, Viana do Castelo, Açores, Vila Real, Braga, Lisboa, Porto, Portalegre, Setúbal, Castelo Branco).

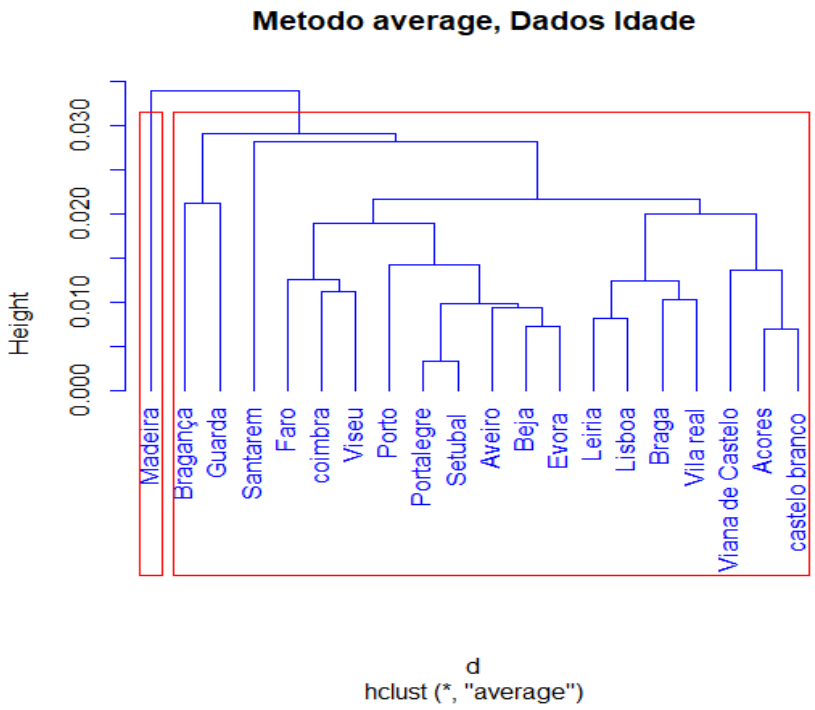

Figura 9: Dendrograma pelo método da média sobre a característica Idade

Assim, do dendrograma da Figura 9, diremos que a distribuição das idades das pessoas que passaram a residir na Madeira é menos similar à distribuição das idades das pessoas que passaram a residir nos restantes distritos (ou ilha dos Açores) os quais formam o segundo *cluster* de distritos e, portanto, são mais homogéneos entre si.

b. Análise de dados da característica *Género* por distrito

A análise do dendrograma obtido pelo método da média sobre a característica Género (Figura 10) apresenta dois grandes grupos. O primeiro grupo é formado por 6 distritos, nomeadamente, Beja, Vila Real, Viana do Castelo, Bragança, Acores, Braga; e o segundo grupo é formado pelos distritos de Setúbal, Santarém, Madeira, Lisboa, Guarda, Coimbra, Faro, Portalegre, Évora, Leiria, Viseu, Porto, Aveiro, e Castelo Branco.

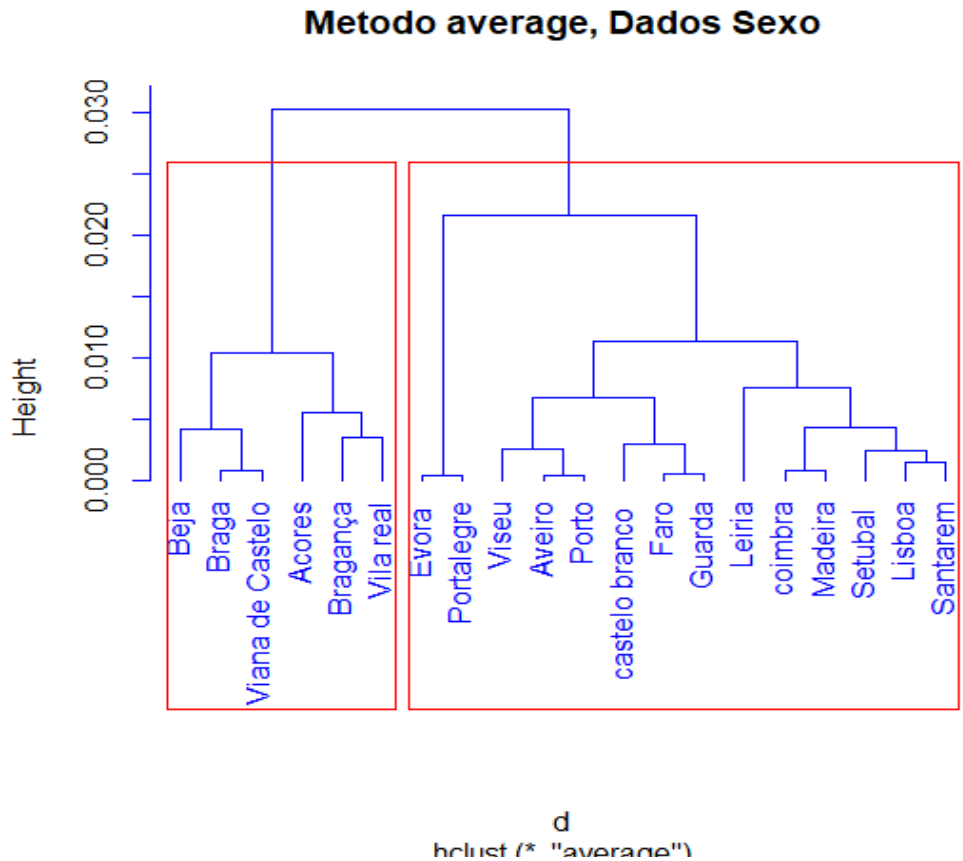

Figura 10: Dendrograma pelo método da média sobre a característica Género

Pelo dendrograma da Figura 10, existem dois grupos de distritos mais homogéneos entre si relativamente à distribuição por género das pessoas que mudaram de residência em Portugal entre 2005 e 2011.

c. Análise de dados da característica *Situação do trabalho* por distrito

Relativamente à situação do trabalho por distrito, da análise do dendrograma da Figura 11 diremos que sobressaem dois grandes grupos. O primeiro grupo é formado pelos distritos de Beja, Braga, Faro, Santarém, Coimbra, Évora, Leiria, Acores, Lisboa, Porto, Madeira, Aveiro e Setúbal; e o segundo grupo é formado pelos distritos de Guarda, Vila Real, Castelo Branco, Bragança, Viana do Castelo, Portalegre e Viseu.

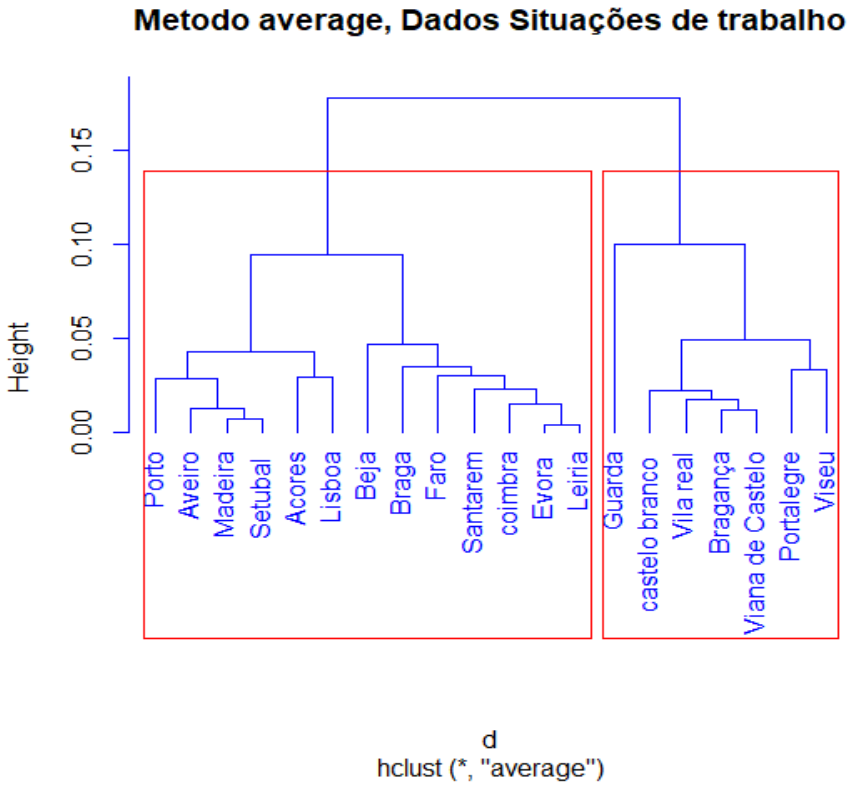

Figura 11: Dendrograma pelo método da média sobre a característica Situação do trabalho

Portanto, de acordo com a Figura 11, existem dois grandes grupos de distritos relativamente bem diferenciados em termos da distribuição da situação de trabalho das pessoas que passaram a residir nesses grupos de distritos. Curiosamente, destaca-se que a distribuição da situação profissional das pessoas que passaram a residir em Lisboa é mais similar à das pessoas que passaram a residir nos Açores (distritos mais homogéneos).

d. Análise de dados da característica *Habilitação literária* por distrito

Analisando o dendrograma correspondente ao método da média para o conjunto de dados relativo às habilitações literárias (Figura 12) conclui-se que existem dois grupos distintos de indivíduos. O primeiro grupo é formado pelos distritos da Guarda, Vila Real, Castelo Branco, Bragança, Viana do Castelo e Viseu; o segundo grupo é formado pelos

distritos de Portalegre, Beja, Braga, Faro, Santarém, Coimbra, Évora, Leiria, Acores, Lisboa, Porto, Madeira, Aveiro e Setúbal.

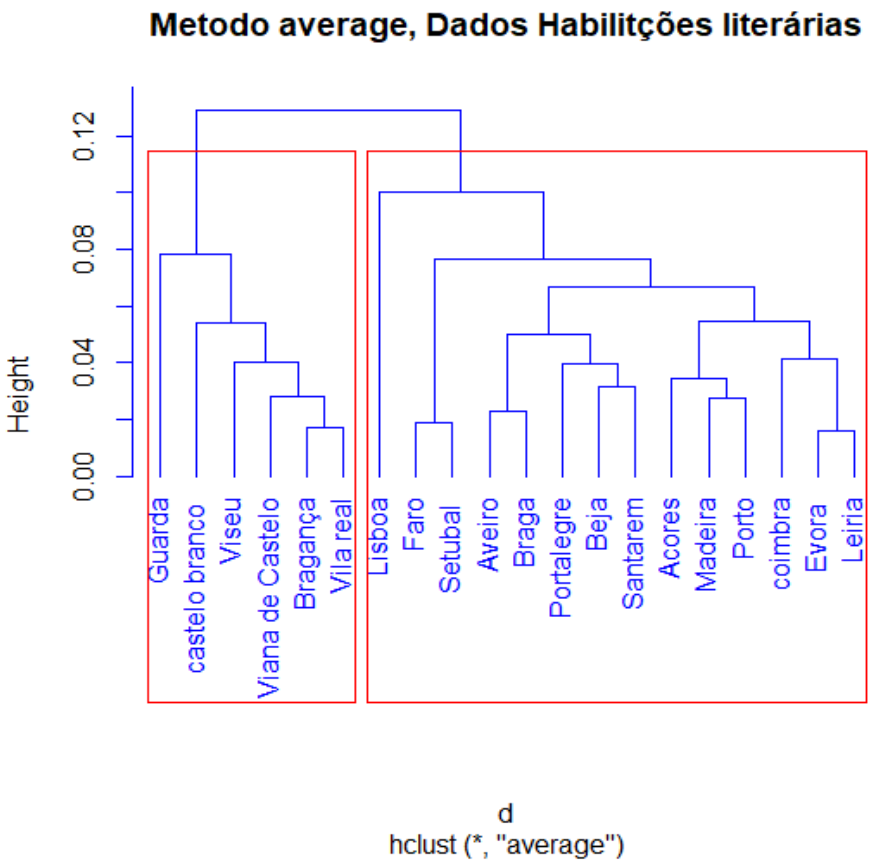

Figura 12: Dendrograma pelo método da média sobre a característica Habilitação literária

A Figura 12 indica que aqueles dois grupos identificados são constituídos, cada um, por distritos mais homogéneos entre si relativamente à distribuição das qualificações académicas pelas pessoas que passaram a residir nos distritos pertencentes a esse mesmo *cluster*. Curiosamente, Aveiro e Braga são distritos que apresentam maiores similaridades assim como Porto é mais similar com as regiões autónomas da Madeira e dos Açores em termos de captação de novos residentes tendo em conta as suas habilitações literárias.

e. Análise de dados nas quatro características em conjunto por distrito

Considerando uma análise com as 20 variáveis em conjunto, a análise do dendrograma resultante do método da média (Figura 13) sugere a existência de dois grandes *clusters*. O primeiro grupo é formado pelos distritos de Guarda, Vila Real, Castelo Branco, Bragança, Viana do Castelo, Portalegre e Viseu e o segundo grupo é formado pelos distritos Beja, Braga, Faro, Santarém, Coimbra, Évora, Leiria, Açores, Lisboa, Porto, Madeira, Aveiro e Setúbal.

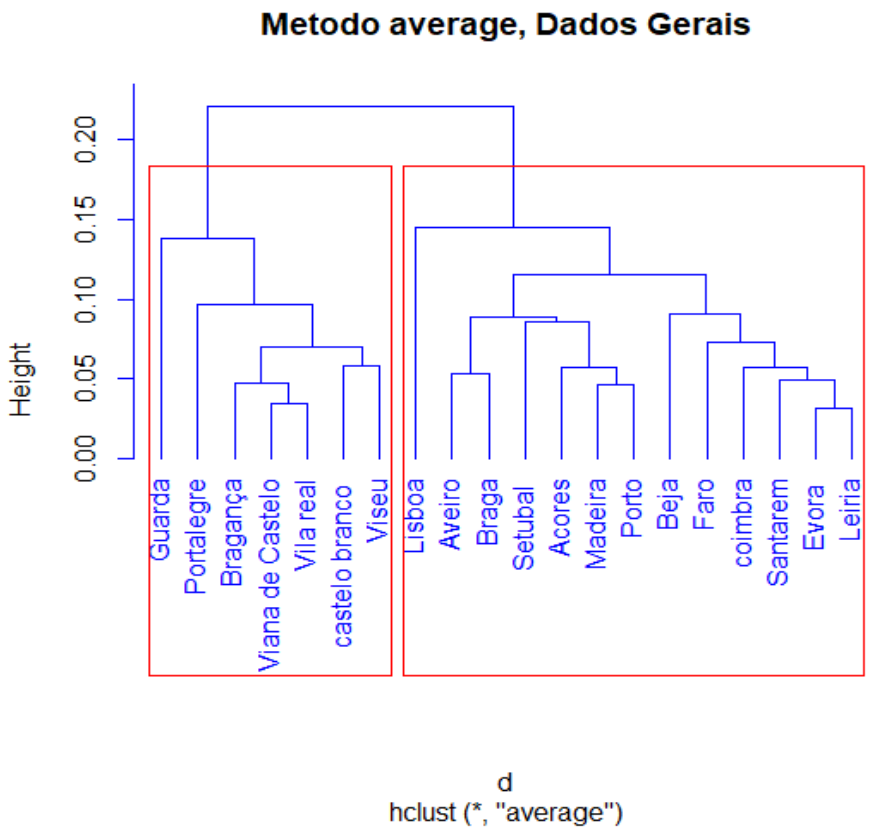

Figura 13:Dendrograma pelo método da média sobre as quatro características em conjunto.

Baseado na Figura 13, podemos deduzir que a movimentação das pessoas que vieram de outro sítio para residir nos distritos assinalados dentro de cada *cluster* apresenta maiores similaridades em termos da distribuição das idades, género, situação de trabalho e habilitação literária.

f. Análise dos dados usando método divisivo DIANA

Usando o DIANA, dois *clusters* distintos de distritos são encontrados sob cada característica, sendo que os valores do coeficiente divisivo são relativamente altos (Tabela 11), com valores próximos de 1. Tal significa que os *clusters* identificados apresentam boa qualidade.

Característica	***Clusters***	**Coeficiente divisivo**
Idade	C_1 = {Açores, Aveiro, Beja, Braga, Bragança, Castelo Branco, Coimbra, Évora, Faro, Guarda, Leiria, Lisboa, Portalegre, Porto, Santarém, Setúbal, Viana do Castelo, Viseu, Vila Real}; C_2 = {Madeira}	0,76
Género	C_1 = {Açores, Beja, Braga, Bragança, Viana do Castelo, Vila Real}; C_2 = {Aveiro, Castelo Branco, Coimbra, Évora, Faro, Guarda, Leiria, Lisboa, Madeira, Portalegre, Porto, Santarém, Setúbal ,Viseu}	0,96
Situação de trabalho	C_1 = {Açores, Beja, Braga, Aveiro, Coimbra, Évora, Faro, Leiria, Lisboa, Madeira, Porto, Santarém, Setúbal}; C_2={Bragança, Castelo Branco, Guarda, Portalegre, Viana do Castelo, Vila Real, Viseu}	0,92
Habilitação literária	C_1 = {Açores, Beja, Braga, Aveiro, Coimbra, Évora, Faro, Leiria, Lisboa, Madeira, Portalegre, Porto, Santarém, Setúbal}; C_2 = {Bragança, Castelo Branco, Guarda, Viana do Castelo, Vila Real, Viseu}	0,84
Dados gerais	C_1 = {Acores, Beja, Braga, Aveiro, Coimbra, Évora, Faro, Leiria, Lisboa, Madeira, Porto, Santarém e Setúbal}; C_2 = {Bragança, Castelo Branco, Guarda, Portalegre, Viana do Castelo, Vila Real, Viseu}	0,82

Tabela 11: Tabela resumo dos clusters obtidos aplicando o método divisivo DIANA

Comparando os grupos identificados pelo método aglomerativo optado (método da média) com os obtidos pelo método divisivo, constatamos que são iguais os conjuntos de *clusters* identificados pelos dois tipos de métodos hierárquicos.

3.3. Métodos de partição

Iremos considerar dois métodos de partição: método de k-médias e o método de k-medóides.

- **Método de k-médias**

Uma vez que neste método necessitamos, à partida, do número *k* de *clusters* homogéneos a considerar, analisaremos, em primeiro lugar, a variação da soma de quadrados dentro dos *clusters* em função do valor de *k*, e escolheremos o valor de *k* de acordo com a Regra do Cotovelo (Figura 14).

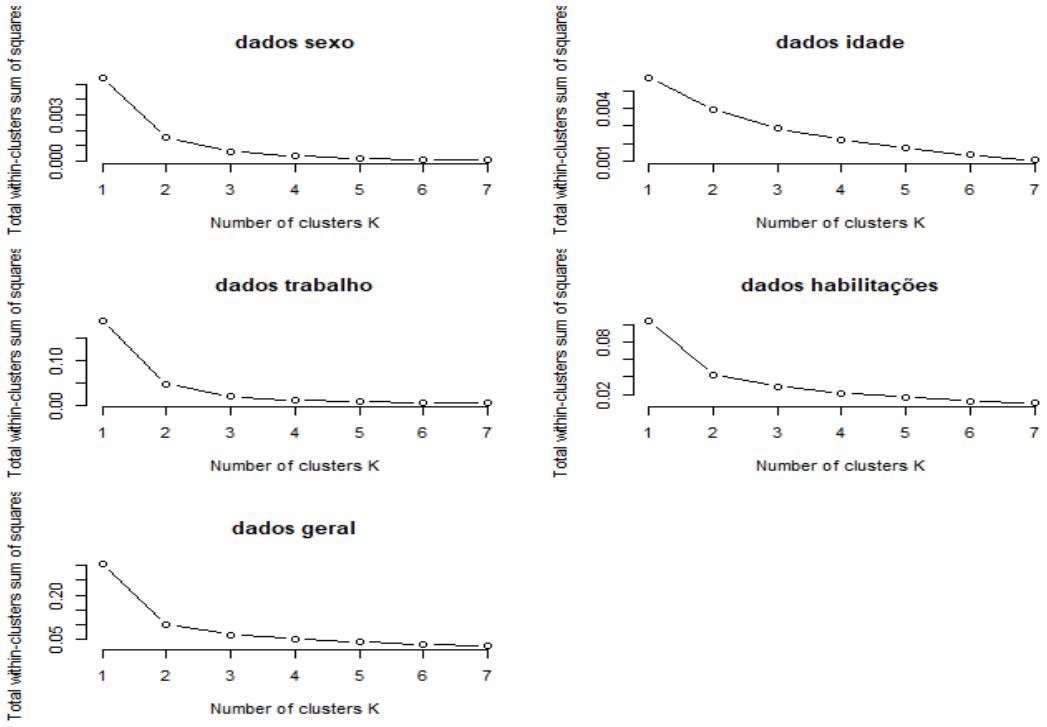

Figura 14: Gráficos para investigar o número de clusters a considerar no método k-médias

Da Figura 14 é visível que a soma de quadrados intra *cluster* (eixos dos yy nos gráficos) tende a diminuir com o aumento sucessivo do número de *clusters*. Usando a Regra do Cotovelo, um número ótimo de *clusters* é considerado como aquele que reporta um " pico" na linha que representa o gráfico da soma de quadrados intra-*clusters* em função do número de *clusters*. A partir destes gráficos (Figura 14) podemos dizer que tomando mais de 2 clusters a diferença observada na soma de quadrado dentro do *clusters* não é substancial. Consequentemente, podemos dizer que será razoável considerar $k = 2$ como o número ótimo de c*lusters* em todas as situações.

A título de curiosidade, para a característica Género, de dimensão 2, representamos na Figura 15 o gráfico de dispersão da distribuição de residentes masculinos e femininos que em 2011 residiam nos 20 distritos indicados mas que em 2005 residiam noutro distrito. Os dois pontos a azul indicam os dois centróides dos dois grupos de distritos identificados pelo método das k-médias. As linhas verticais e horizontais assinalam o valor de referência 0,5 (tantos masculinos como femininos). Observa-se que no *cluster* com menor número de distritos (assinalados a vermelho) há maior tendência para a percentagem de residentes que mudança de distrito de residência terem mais homens do que mulheres; em particular, fazem parte desse grupo, os distritos de Viana do Castelo, Braga, Beja e a região autónoma dos Açores. Em oposição, nos distritos de Évora e de Portalegre, pertencentes ao outro *cluster*, existe um maior número de mulheres do que de homens que mudaram de residência para esses distritos.

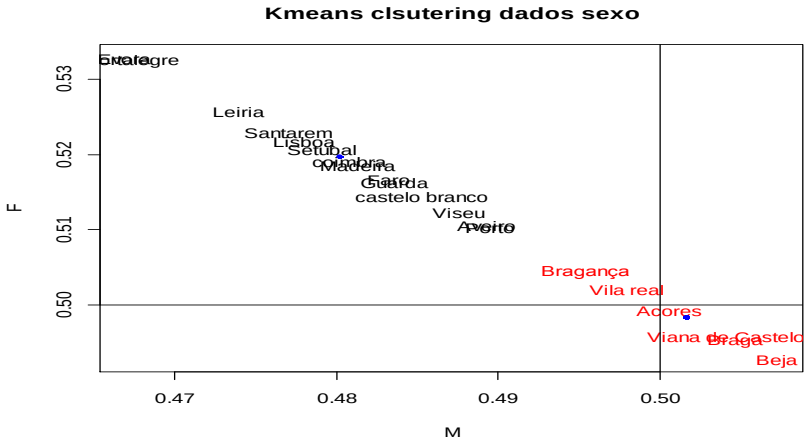

Figura 15:Gráfico de dispersão para a característica género, diferenciando os clusters encontrados pelo método k-médias com k=2

- **Método de partição de k-medóides**

Nesta fase são apresentados os resultados obtidos do algoritmo de k-medóides usando a distância euclidiana. Para determinar o número de *clusters*, uma ferramenta útil é o índice de silhueta cujo resultado média das médias é apresentado na Tabela 12.

Característica Género								
k	2	3	4	5	6	7	8	9
Médias das médias	0,62	0,59	0,52	0,54	0,58	0,57	0,58	0,54
Característica Idade								
k	2	3	4	5	6	7	8	9
Médias das medias	0,24	0,29	0,24	0,21	0,22	0,21	0,23	0,19
Característica Situação do trabalho								
k	2	3	4	5	6	7	8	9
M**é**dias das m**é**dias	0,63	0,62	0,59	0,54	0,45	0,39	0,35	0,26
Característica Habilitação literárias								
k	2	3	4	5	6	7	8	9
M**é**dias das m**é**dias	0,49	0,35	0,28	0,32	0,31	0,35	034	0,33
Dados gerais								
k	2	3	4	5	6	7	8	9
Médias das médias	0,53	0,38	0,29	0,28	0,27	0,26	0,25	0,24

Tabela 12:Resultados do método k-medóides através do método das silhuetas

Tendo em conta os valores de média das médias para o índice de silhueta encontrados na Tabela 12, conclui-se que o método de k-medóides dará uma formação de *clusters* com

- uma estrutura razoável para a característica género com o número *k* a variar de 2 até 9;
- sem nenhuma estrutura substancial para a caraterística idade com o número *k* a variar de 2 a 9;
- uma estrutura razoável para a característica situação de trabalho com *k* a variar de 2 a 5, mas uma estrutura fraca com *k* a variar de 6 a 9;

- uma estrutura fraca e pode ser artificial para a característica habilitação literária com o número k a variar de 2 até 9;
- uma estrutura razoável para o conjunto de dados gerais quando k é 2, uma estrutura fraca quando k varia de 3 a 7, e uma formação de *cluster* sem nenhuma estrutura substancial quando k varia de 8 a 9.

Assim sendo, o valor de k que produz melhores estruturas de formação de clusters é $k = 2$. Para a escolha dos dois medóides considerou-se o algoritmo implementado no comando `pam` do pacote "clValid" do R. As movimentações das pessoas que entram num distrito de Portugal, tendo em conta os dados do sexo, idade, situação do trabalho e habilitação literária, indicam um primeiro grupo composto pelos distritos de Açores, Aveiro, Beja, Braga, Coimbra, Évora, Faro, Leiria, Lisboa, Madeira, Porto, Santarém e Setúbal, e um segundo grupo composto pelo distritos de Bragança, Castelo Branco, Guarda, Portalegre, Viana do Castelo, Vila Real e Viseu. O gráfico de silhueta está representado na Figura 16.

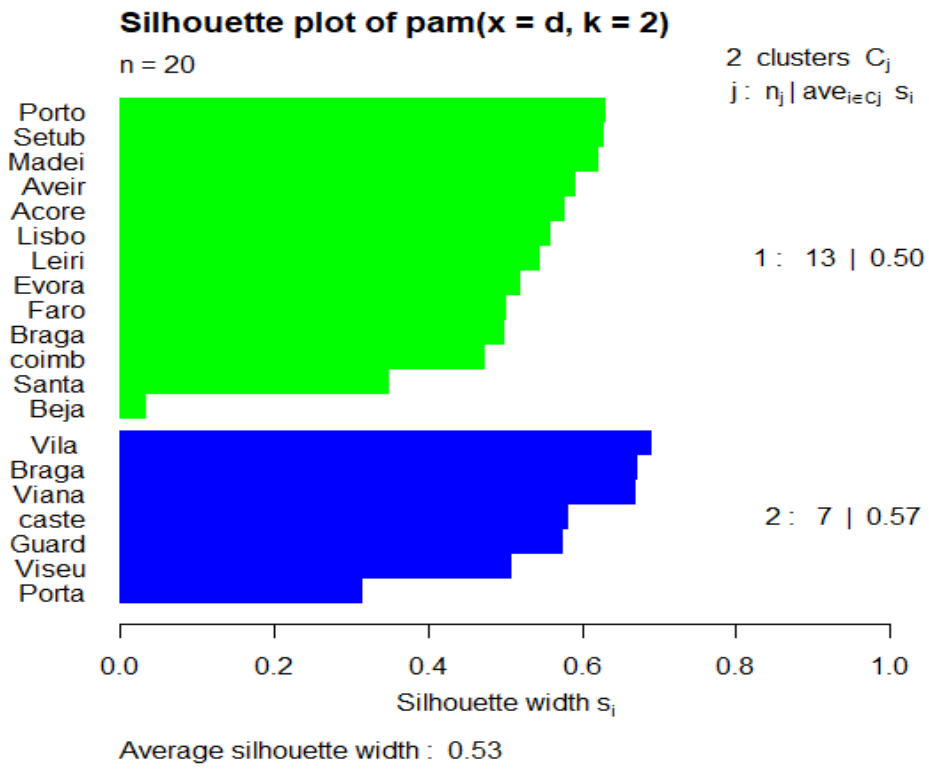

Figura 16:Gráfico da aplicação do método silhueta com k=2 para os dados gerais

3.4. Validação interna e validação de estabilidade

Para além do coeficiente correlação cofenética, técnicas de validações interna e de estabilidade podem ser utilizadas para investigar os valores ótimos dos métodos hierárquicos e não hierárquicos. No nosso caso, vamos aplicá-las para os dados relativos à idade, situação do trabalho, habilitação literária e dados gerais. Para os dados relativos ao sexo, uma vez que é 2-dimensional, podemos observar direitamente num referencial 2-dimensional as movimentações, em termos de frequências, de mulheres e de homens que passaram a residir nos distritos assinalados (por exemplo, como na Figura 15, com a visualização dos grupos dos distritos por cores identificados pelo método k-médias). Assim, passamos de seguida a analisar as validações interna e de estabilidade por característica. Como método hierárquico consideramos o método aglomerativo com o critério das médias.

Começamos pela característica idade.

Validação interna para a característica Idade

Métodos	Índices	k	Valor ótimo
Hierárquico	Conectividade	2	3,0290
Hierárquico	Dunn	2	0,5340
Hierárquico	Silhueta	2	0,0107

Tabela 13:Validação interna dos dados idade assinalando o número de clusters k onde o valor ótimo da medida foi atingido e qual o método para qual tal resultou.

Pelos resultados da Tabela 13, os valores de ótimalidade de validação interna indicam o método hierárquico com dois *clusters* que produz melhores resultados. O valor de *k* ótimo pode ser obtido por visualização gráfica de cada uma das medidas de validação (Figura 17):

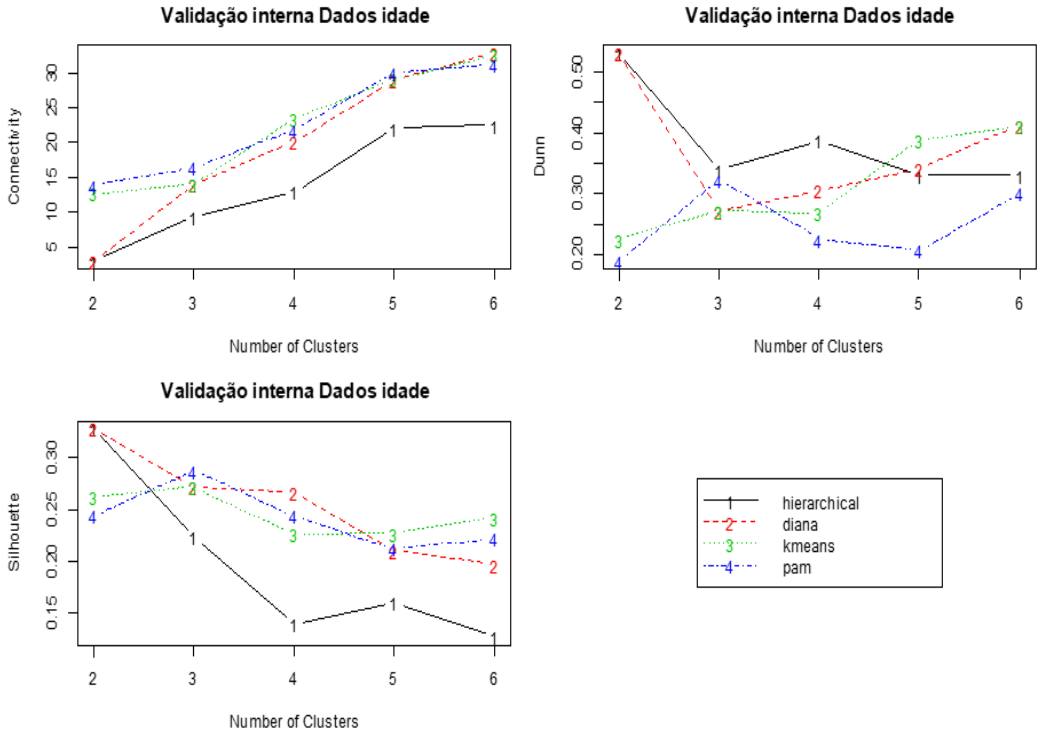

Figura 17:Gráfico validação interna dados idade

A Figura 17 permite analisar a otimalidade de cada medida de validação (Conectividade, Índice de Dunn e o índice de Silhueta) quando *k* varia de 2 a 6 *clusters*. Os gráficos demonstram que os agrupamentos hierárquicos (aglomerativo ou divisivo) superam os outros dois algoritmos de agrupamento em cada medida de validação. Para o agrupamento hierárquico, o número ótimo de *clusters* é claramente, segundo as medidas conectividade, Dunn e Silhueta, o número dois. Uma análise mais detalhada dos valores desses índices, quando k=2, demonstram que é no método aglomerativo que se atingem os melhores resultados para os índices (Tabela 13); portanto, o agrupamento hierárquico aglomerativo com 2 *clusters* tem a melhor pontuação.

Analisemos agora a estabilidade. Pelos resultados da Tabela 14, para as quatro medidas, o valor de otimalidade da validação indica que a estabilidade do agrupamento é atingida para o método hierárquico (aglomerativo) e o k-medóides, com dois *clusters*, de acordo com as medidas *APN* e *ADM*, e com seis *clusters* de acordo com as medidas *AD* e *FOM*. Isto é, aqueles dois métodos indicam as suas estabilidades com melhor resultado.

Validação da estabilidade para a característica Idade

Método	Índices	k	Valor ótimo
Hierárquico	*APN*	2	0,0358
k-medóides	*AD*	6	0,0105
Hierárquico	*ADM*	2	0,0012
k-medóides	*FOM*	6	0,0055

Tabela 14:Validação da estabilidade dos dados idade assinalando o número de clusters k onde o valor ótimo do índice foi atingido e qual o método para qual tal resultou

O resultado da Tabela 14 pode ser visualizado, de forma menos precisa, com o gráfico de cada uma das medidas de validação (Figura 18)

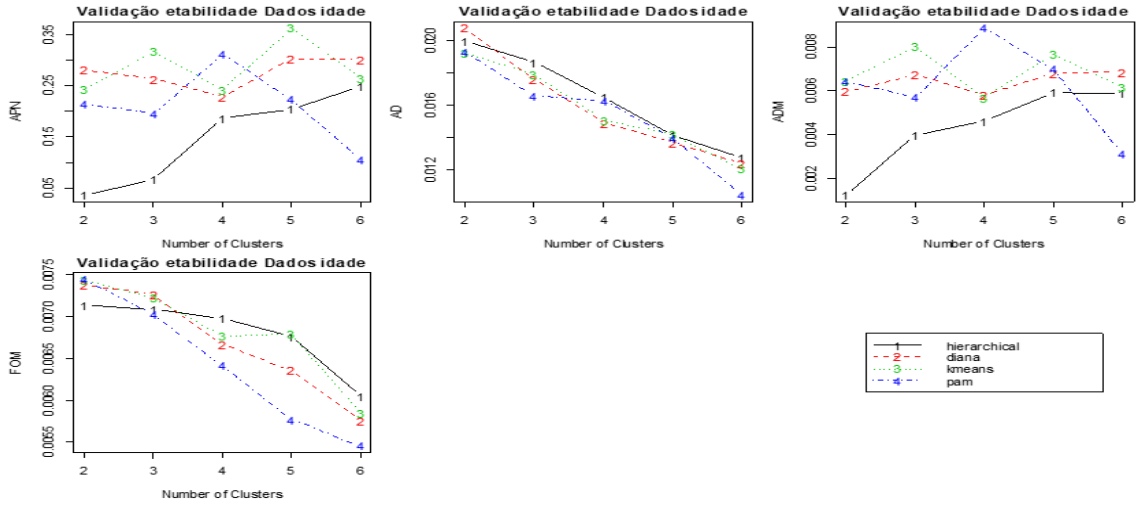

Figura 18:Gráfico de validação da estabilidade para os dados idade

Dos graficos das Figura 18 observa-se que a medida APN mostra uma tendência interessante para o método k-medóides, na medida em que aumenta de três a quatro *clusters*, e de cinco a seis *clusters* decrescer sendo que para os métodos aglomerativos e divisivos aumenta. Contudo, o método hierárquico (aglomerativo) com dois *clusters* apresenta o melhor resultado. As medidas *AD* e *FOM* tendem a diminuir à medida que o número de *clusters*

aumenta. Aqui o método k-medóides com seis *clusters* tem o melhor resultado geral. Para a medida *ADM*, o hierárquico com dois *clusters* tem novamente a melhor pontuação.

Vejamos agora para as outras características.

Validação interna para a característica Situação do trabalho

Método	Índice	k	Valores ótimos
Hierárquico	Conectividade	2	4,7012
Hierárquico	Dunn	4	0,4661
Hierárquico	Silhueta	2	0,6293

Tabela 15:Validação interna dos dados situação do trabalho assinalando o número de clusters k onde o valor ótimo do índice foi atingido e qual o método para qual tal resultou

Para a característica situação do trabalho, pelos resultados da Tabela 15, os valores de otimalidade da validação interna indicam o método hierárquico com dois *clusters,* tendo em conta os índices de conectividade e de silhueta, e com quatro *clusters* tendo em conta o índice Dunn. Isto é, os métodos hierárquicos nos dados de situação do trabalho indicam os melhores resultados. Os gráficos para cada uma das medidas de validação está na Figura 19.

Os gráficos da Figura 19 ilustram *k* a variar de dois a seis *clusters* segundo os índices de conectividade, Dunn e silhueta. Da visualização destes gráficos resulta que o *k* óptimo será 2 de acordo com as medidas conectividade e silhueta e 4 de acordo com a medida de Dumm independentemente do método de agrupamento. Uma analise mais precisa dos valores (taebla 15) resulta que para o agrupamento hierárquico, o número ótimo de *cluster* é segundo as medidas de conectividade e de silhueta, o número dois e, segundo o índice de Dunn, o número quatro; portanto, o agrupamento hierárquico tem a melhor pontuação. Relativamente à estabilidade (Tabela 16), para as quatro medidas, o valor de otimalidade da validação obtido indica a estabilidade do *clusters* no método hierárquico e no método k-medóides, onde duas medidas (*APN* e *ADM*) sugerem dois *clusters* e outras duas medidas (*AD* e *FOM*) sugerem seis *clusters*. Isto é, aqueles dois métodos indicam as suas estabilidades com melhor resultado.

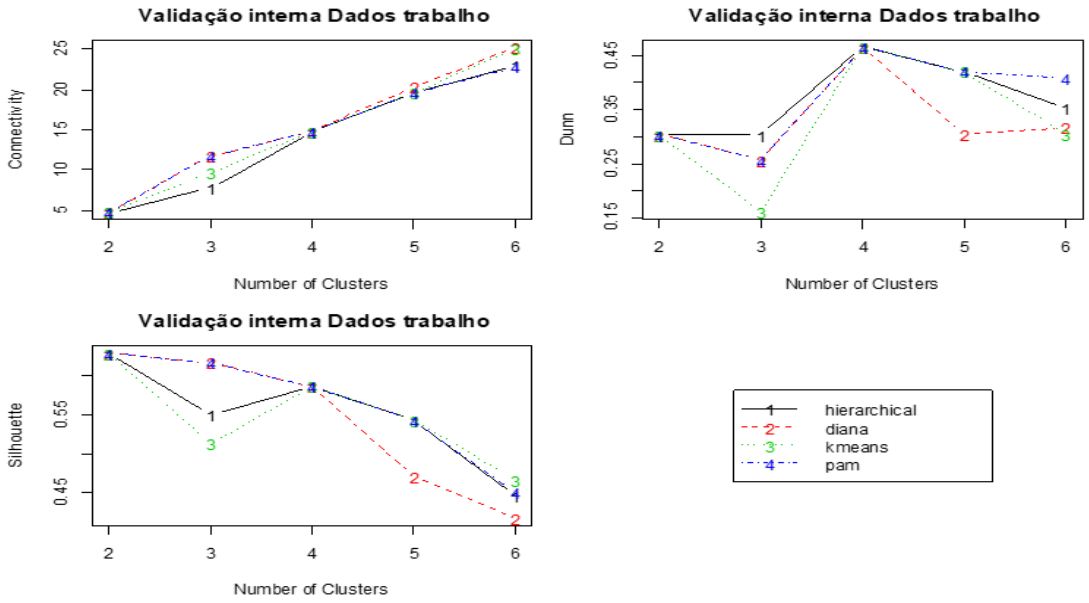

Figura 19:Validação interna para os dados situação do trabalho

Validação de estabilidade para a característica Situação do trabalho

Método	Índices	K	Valores ótimos
Hierárquico	*APN*	2 e 4	0,0000
k-médias	*APN*	4	0,0000
k-medóides	*APN*	2 até 4	0,0000
k-medóides	*AD*	6	0,0209
Hierárquico	*ADM*	2	0,0000
k-medóides	*ADM*	2 até 4	0,0000
k-medóides	*FOM*	6	0,0131

Tabela 16:Validação da estabilidade para os dados situação do trabalho

Na Figura 20 são ilustrados os gráficos de cada uma das medidas de estabilidade para a situação de trabalho.

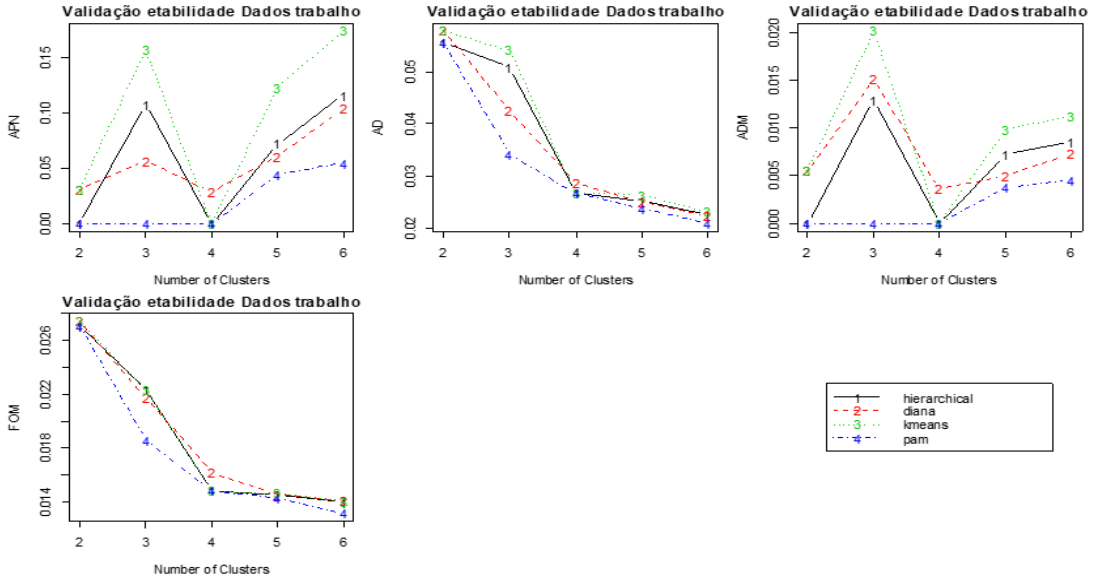

Figura 20:Validação da estabilidade para os dados situação do trabalho

Conforme a Figura 20, a medida *APN* mostra uma tendência interessante, na medida em que aumenta inicialmente de dois a três *clusters*, diminui para quatro *clusters* e posteriormente aumenta. Os métodos hierárquicos (aglomerativo e o divisivo) com dois e o quatro *clusters* têm o melhor resultado; o método k-médias tem o melhor resultado com quatro clusters e, por outro lado, o k-medóides tenha o melhor resulado de estabilidade com o número de clusters a variar de dois a quatro. Por seu turno, as medidas *AD* e *FOM* tendem a diminuir à medida que o número de *clusters* aumenta. Aqui o k-medóides com seis *clusters* tem o melhor resultado geral, embora os outros algoritmos tenham pontuações semelhantes. Para a medida *ADM*, os métodos hierárquicos (aglomerativo e divisivo) indicam o valor ótimo com dois *clusters* , contudo o k-medóides indica o valor ótimo com clusters a variar de 2 a quatro, isto é, o k-medóides apresenta o melhor resultado de estabilidades.

Finalmente, para a característica habilitação literária, os valores de otimalidade da validação interna com base nos índices de conectividade e silhueta correspondem ao método hierárquico com dois *clusters* enquanto que o método divisivo com seis *clusters* é sugerido como o melhor em termos de validação interna usando o índice de Dunn (Tabela 17).

Validação interna para a característica Habilitação literária

Método	Índice	k	Valores ótimos
Hierárquico	Conectividade	2	5,5206
Diana	Conectividade	2	5,5206
k-médias	Conectividade	2	5,5206
k-medóides	Conectividade	2	5,5206
Diana	Dunn	6	0,6759
k-médias	Dunn	6	0,6759
Hierárquico	Silhueta	2	0,4887
Diana	Silhueta	2	0,4887
k-médias	Silhueta	2	0,4887
k-medóides	Silhueta	2	0,4887

Tabela 17:Validação interna para os dados habilitação literária

Portanto, pelos resultados da Tabela 17, os valores de otimalidade de validação interna indica o método hierárquico com dois *clusters*, com base nos índices conectividade e silhueta, e o método divisivo com seis *clusters* se o índice de Dunn for considerado. Isto é, os métodos hierárquicos e divisivos nos dados de habilitação literária indicam os melhores resultados. Os gráficos desses três índices como medidas de validação, para *k* a variar de dois a seis *clusters* encontram-se na Figura 21.

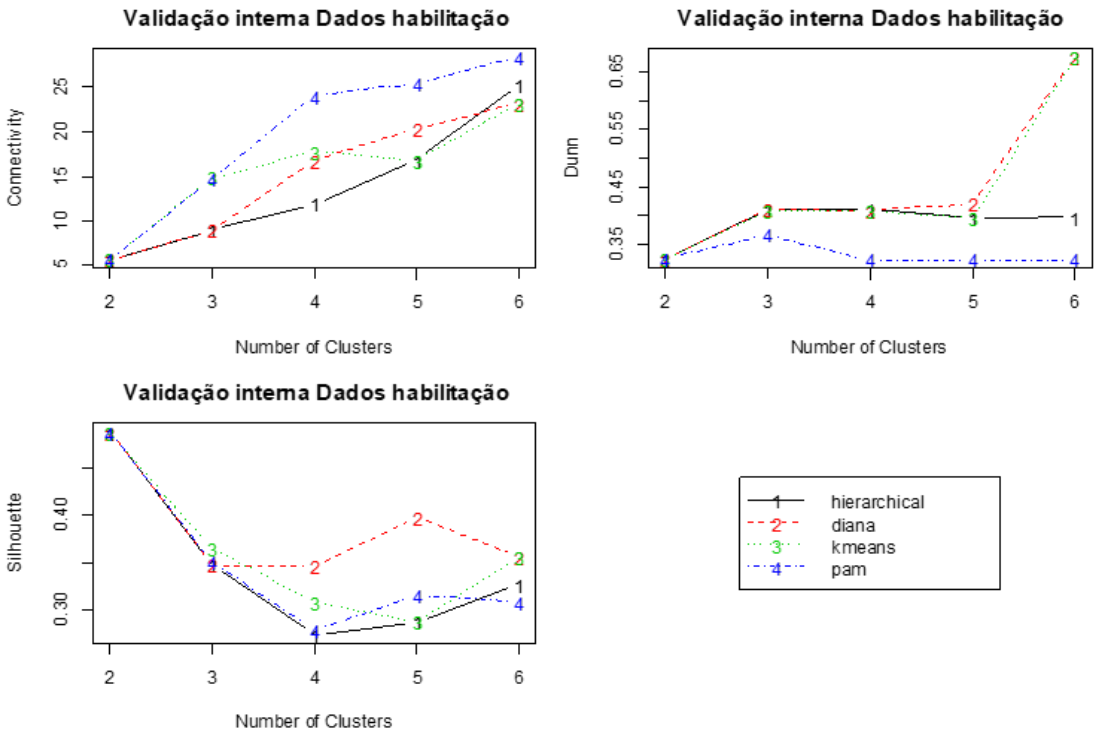

Figura 21:Validação interna dos dados habilitação literária

Na Figura 21 o índice de conectividade apresenta os valores ótimos com dois *clusters* para todos métodos, ou seja, para o método hierárquico (aglomerativo), diana, k-médias e k-medóides. Por outro lado, o índice de Dunn, indica o melhor resultado ótimo com seis *clusters* para o método diana e k-médias, enquanto o índice de silhueta mostra a melhor pontuação para todos os métodos e com $k = 2$.

Validação de estabilidade para a característica Habilitação literária

Métodos	Índices	k	Valores ótimos
Hierárquico	*APN*	2	0,0083
Diana	*AD*	6	0,0299
k-medóides	*ADM*	2	0,0037
Diana	*FOM*	6	0,0099

Tabela 18:Validação de estabilidade para dados habilitação literária

Relativamente à estabilidade, pelo resultado da Tabela 18, para as quatros medidas, o valor de otimalidade da validação de estabilidade surge para o método hierárquico (Aglomerativo), o k-medóides e o Diana sendo que duas medidas (*APN* e *ADM*) assinalam dois *clusters*, e as outras duas medidas (*AD* e *FOM*) assinalam seis *clusters*. Isto é, os três métodos indicam as suas estabilidades com melhor resultado. Os gráficos destas medidas podem ser visualizados na Figura 22.

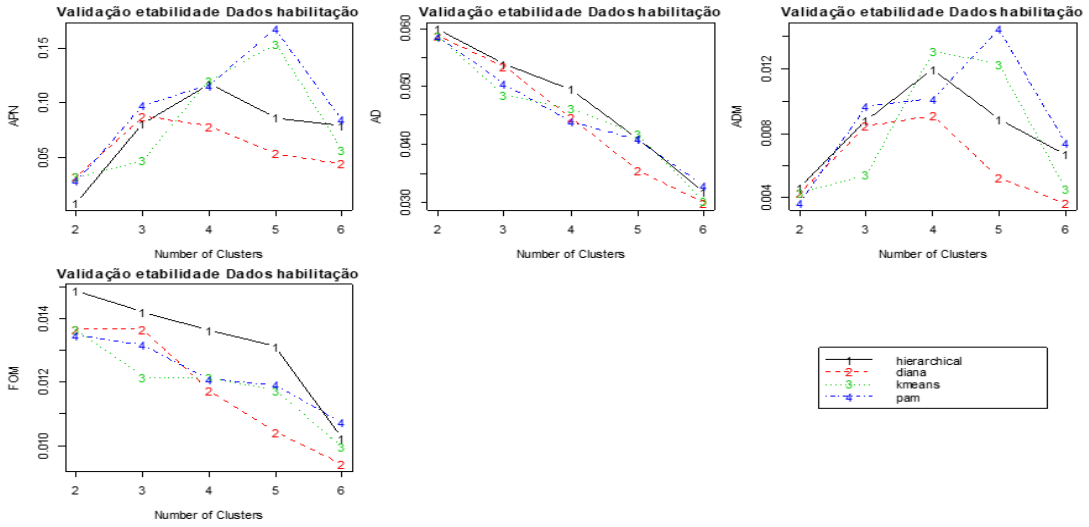

Figura 22:Validação estabilidade para dados habilitação literária

Na Figura 22 observa-se que a medida *APN* mostra uma tendência interessante, na medida em que aumenta inicialmente de dois a três *clusters*, e diminui no quatro *clusters* para o diana mas aumenta para os restantes três metodos. Do grafico de *APN*, o hierárquico com dois *clusters* tem o melhor resultado. As medidas *AD* e *FOM* tendem a diminuir à medida que o número de *clusters* aumenta. Aqui o Diana com seis *clusters* tem o melhor resultado geral, embora os outros algoritmos tenham pontuações semelhantes. Para a medida *ADM*, o k-medóides com dois *clusters* tem a melhor pontuação.

Analisemos agora as 20 variáveis em conjunto (dados gerais).

Validação interna para os dados gerais

Método	Índice	k	Valores ótimos
Hierárquico	Conectividade	2	5,0333
Diana	Conectividade	2	5,0333
k-médias	Conectividade	2	5,0333
k-medóides	Conectividade	2	5,0333
Hierárquico	Dunn	5	0,6760
Hierárquico	Silhueta	2	0,5275
Diana	Silhueta	2	0,5275
k-médias	Silhueta	2	0,5275
k-medóides	Silhueta	2	0,5275

Tabela 19:Validação interna para os dados gerais

Pelo resultado da Tabela 19, os valores de otimalidade da validação interna optam pelo método hierárquico com dois *clusters*, com base nos índices conectividade e silhueta, e com cinco *clusters* com base no índice de Dunn. Isto é, os métodos hierárquicos para os dados gerais indicam os melhores resultados. Na Figura 23 estão os gráficos associados a aqueles três índices.

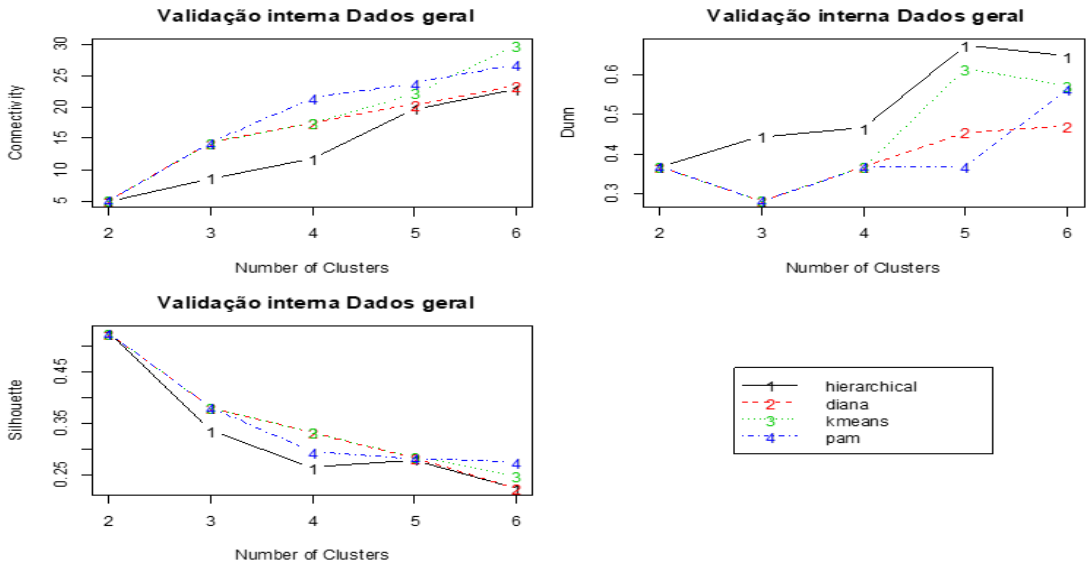

Figura 23:Validação interna para os dados gerais

Conforme a Figura 23, observamos que os índices de conectividade e de silhueta indicam as melhores pontuações com dois *clusters* para todos os métodos como o método hierárquico, diana, k-médias, e do k-medóides . Para o agrupamento hierárquico, segundo as medidas Dunn, o número ótimo de *clusters* é dois.

Validação de estabilidade para os dados gerais

Método	Índices	K	Valores ótimos
k-medóides	*APN*	2	0,0043
k-medóides	*AD*	6	0,0483
k-medóides	*ADM*	2	0,0011
k-medóides	*FOM*	6	0,0107

Tabela 20:Validação de estabilidade dados geral

Pela Tabela 20, relativa às quatro medidas de estabilidade, o valor de otimalidade da validação indica o método k-medóides, onde duas medidas (*APN* e *ADM*) optam por dois *clusters*, e as outras duas (*AD* e *FOM*) optam por seis *clusters*. Isto é, o método de k-medóides (PAM) apresenta as suas estabilidades com melhor resultado. Para visualizar os gráficos destas quatro medidas temos a Figura 24.

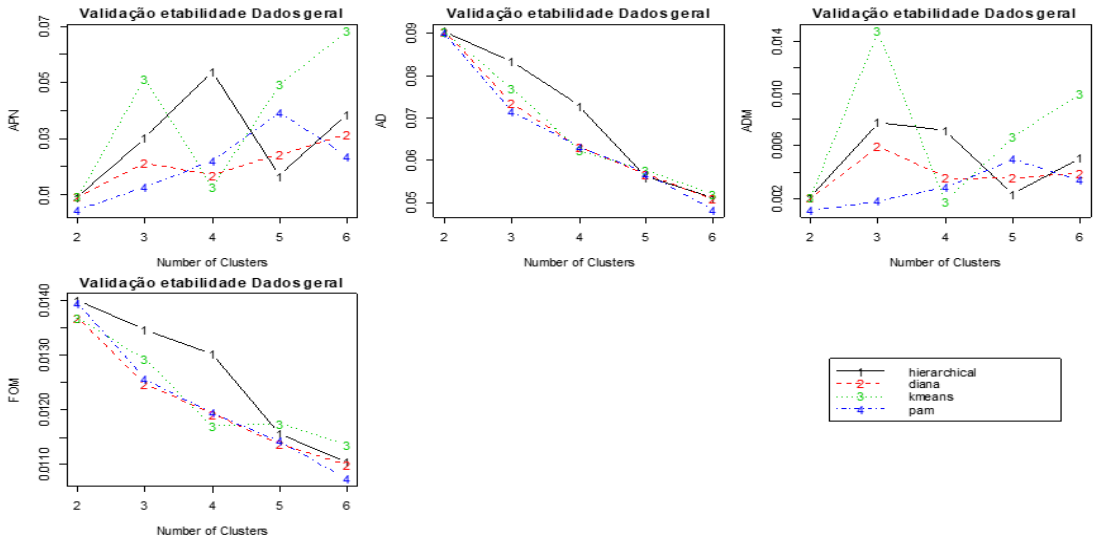

Figura 24:Validação da estabilidade para os dados gerais

Na Figura 24, a medida *APN* mostra uma tendência interessante na medida em que aumenta inicialmente de dois a três *clusters* para os quatro métodos mas depois varia de maneira diferente entre os 4 metodos. O método k-medóides (PAM) com dois *clusters* tem o melhor resultado. As medidas *AD* e *FOM* tendem a diminuir à medida que o número de *cluster* aumenta. Aqui o PAM com seis *clusters* tem o melhor resultado geral, embora os outros algoritmos tenham pontuações semelhantes. Para a medida da *ADM*, o PAM com dois *clusters* tem a melhor pontuação, embora os outros métodos superem o PAM para um número maior de *clusters* (4 , 5 ou 6 *clusters*).

Capítulo 4

Considerações finais

Neste capítulo são apresentadas algumas considerações finais, sumariados resultados analisados no capítulo anterior e algumas sugestões para futuro.

4.1. Conclusão

O objetivo desta dissertação foi descrever métodos de agrupamento, de indivíduos ou de variáveis, e ainda índices que permitem identificar qual é a melhor técnica de agrupamento segundo algum critério de otimalidade. Estas técnicas e índices foram aplicados a uma base de dados reais relativa à movimentação das pessoas que mudaram de residência nos distritos do território do país de Portugal entre o ano 2005 e 2011. Os dados foram analisados segundo quatro características: sexo, idade, situação do trabalho e habilitação literária.

Após aplicar vários critérios de agregação com a medida de distância euclidiana, foi identificada a melhor técnica hierárquica entre métodos aglomerativos através do coeficiente de correlação cofenética. Os resultados do coeficiente correlação cofenética indicam o método das médias (*average*) como o método com melhor desempenho para a base de dados em estudo.

Os valores do coeficiente de correlação cofenética pelo método das médias em cada caso são: dados sexo com valor 0,7898 (79,0%), dados idade com valor 0,7525 (75,3%), dados de situação do trabalho com valor 0,7964 (79,6%), e dados habilitação literária com valor 0,7467 (74,6%). Assim, os dados mais influentes neste agrupamento, no método das médias, são os dados da situação do trabalho pois indica a maior percentagem que é 79,6%. Isto significa uma melhor definição dos grupos quando se analisam as situações de trabalho na movimentação das pessoas que vieram ou entram nos distritos do território de Portugal entre ano 2005 e 2011.

Recorrendo ao método de partição das k-médias, a regra do cotovelo indicou $k = 2$ sobre as distintas características analisadas. Considerando o método k-medóides, o método de silhuetas indicou diferentes números de *clusters* consoante a característica em estudo. Tomando os dados em geral, o índice de silhueta indicou uma boa estrutura do agrupamento com o número de *clusters* $k = 2$.

Para além da coeficiente de correlação cofenética outras medidas de validação foram usadas: validação interna e estabilidade, com o objetivo de avaliar as qualidades e estabilidade dos *clusters* formados por todos os métodos da Análise de *Clusters*, tanto hierárquicos como não hierárquicos estudados.

Relativamente ao conjunto de dados estudado verificou-se, pelos resultados das validações, que a maioria dos métodos hierárquicos com critério de aglomeração da média (*average)* acrescenta as melhores qualidades e estabilidade dos *clusters* quando analisadas as diferentes características separadamente. Agregando toda a informação numa só matriz de dados, verificou-se que tal método hierárquico exibe qualidade, mas não estabilidade dos c*lusters*.

Deste modo conclui-se que, relativamente aos dados estudados, entre os critérios de agregação, o método da média (*average*) indica os melhores desempenhos de qualidade e estabilidade de *clusters* quando comparado com os outros.

4.2. Sugestões

De acordo com as conclusões acima, esta dissertação faz uma comparação de técnicas de aglomeração com o coeficiente de correlação cofenética, métodos de validação interna e validação de estabilidade usando como medida de dissimilaridade a distância euclidiana. Portanto, sugerimos que futuramente se aplique outras medidas de comparação de agrupamentos e outras medidas de validação. E também outras medidas de dissimilaridade, de modo a fazer comparações usando diferentes critérios de agregações e métodos hierárquicos e não hierárquicos. Espera-se, em futuras aplicações de Análise de *Clusters* em outros casos, como medicina, agricultura, engenharia, recorrer às diferentes medidas aqui consideradas com

vista a auxiliar na escolha do método de agrupamento com melhores qualidades para o conjunto de dados em análise.

Bibliografia

Anderberg, M. R. (1973). *Clusters analysis for applications*. Academic Press. NY.

Jain, A. K., Dubes, R. C. (2008).*Clustering*. Wiley.

Berkhin, P. (2002). Survey of Clustering data mining techniques. *Accrue Software*, San Jose, CA, 1–56.

Brock, G., Pihur, V., Datta, S. S., Datta, S. S. (2008). clValid : An R package for cluster validation. *Journal Of Statistical Software*, 25, 1–28.

Everitt, B. S., Landau, S., Leese, M., Stalh, D. (2011).*Clusters analysis* (5th Edition). Wiley.

Government of Timor-Leste. (September ,2015). Population and housing census preliminary results.

Gower, J. C., Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3, 5–48.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 445–492.

Jain, A. K, Murty, M.N., Flynn, P.J. (1999). Data Clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323

Johnson, R., Wichern, D. (2014). *Applied multivariate statistical analysis* (6ª Edição). Pearson Education, Sydney.

Kaufman, L., Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Igarss. Canada.

Kodinariya, Trupti M, and Prashant R Makwana. (2013). Review on determining number of clusters in K-means clustering. *International Journal of Advance Research in Computer*

Science and Management Studies 1(6): 90-94.

Metz, J., Monard, M. C. (2005). Clustering hierárquico : uma metodologia para auxiliar na interpretação dos clusters. *Atas do XXV Congresso da Sociedade Brasileira da Computação*.1170–1173.

Murtagh, F., Legendre, P. (2014). Ward's hierarchical agglomerative *Clustering* method : Which algorithms implement Ward's criterion ? *Journal of Classification*, *31*, 274–295.

Reis, E. (2001). *Estatística multivariada aplicada* (2ª Edição). Lisboa.

Rencher, A. C. (2012). *Methods of multivariate analysis* (3ª Edição). Wiley.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Saraçli, S., Dogan, N., Dogan, I. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, *1*(203), 1–8.

Sharma, S. (1996). *Applied multivariate techniques*. Wiley.

Simar, H. (2007). *Applied multivariate statistical analysis* (2ª Edição). Springer. Berlim.

Sokal, R. R., Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33–40.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680.

Timm, N. H. (2002). *Applied multivariate analysis*. *Transplant Immunology*. Springer-Verlag New York.

Williams, W. T., Lambert, J. M. (1959). Multivariate methods in plant Ecology. I. Association-analysis in plant communities. *J. Ecology*, 47(1), 83–101.

Wunsch II, D. C., Xu, R. (2008). *Clustering*. Wiley.

Anexo A: Conjunto de dados reais analisados

"Distritos"	Género		Situação de trabalho			Idade					Total
	M	F	T1	T2	T3	I1	I2	I3	I4	I5	
Açores	7925	7906	1277	8948	5606	605	2269	5722	5725	1510	15831
Aveiro	20960	21879	3945	22874	16020	1632	5770	14941	15733	4763	42839
Beja	5782	5617	1089	5223	5087	396	1517	3997	4137	1352	11399
Braga	22200	21793	4138	22413	17442	1849	5801	16237	15485	4621	43993
Bragança	4601	4686	694	3672	4921	377	1282	3142	3462	1024	9287
Castelo Branco	5697	6039	949	4800	5987	441	1660	4304	4199	1132	11736
Coimbra	14949	16143	2435	15356	13301	1343	4312	10973	10982	3482	31092
Évora	4984	5688	928	5275	4469	377	1462	3771	3849	1213	10672
Faro	19929	21314	4460	19879	16904	1588	6038	14827	14309	4481	41243
Guarda	4763	5086	779	3338	5732	389	1370	3352	3515	1223	9849
Leiria	14708	16324	2779	15347	12906	1155	4186	11457	10627	3607	31032
Lisboa	129554	141465	21894	158773	90352	9728	36397	100603	94296	29995	271019
Madeira	10008	10785	1896	11332	7565	936	2442	7548	7811	2056	20793
Portalegre	3143	3583	752	2750	3224	263	890	2405	2425	743	6726
Porto	64085	66828	13644	72136	45133	5056	16390	47397	46901	15169	130913
Santarém	17744	19455	3263	17757	16179	1541	5397	13416	12446	4399	37199
Setúbal	44117	47950	8732	49683	33652	3420	12164	32866	33424	10193	92067
Viana do Castelo	7852	7726	1261	6204	8113	564	2057	5789	5622	1546	15578
Vila Real	6463	6517	1185	5033	6762	462	1748	4733	4605	1432	12980
Viseu	10927	11483	1910	9585	10915	860	3317	7849	7933	2451	22410

"Distritos"	Habilitação literária										Total
	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	
Açores	297	76	2299	2300	2554	251	2473	3359	383	1839	15831
Aveiro	919	127	7139	6615	8193	683	6789	6457	839	5078	42839
Beja	197	18	2161	1587	2045	142	2027	1426	189	1607	11399
Braga	714	242	7966	6799	7853	565	6656	6974	983	5241	43993
Bragança	152	14	2418	1255	1409	120	1191	1301	150	1277	9287
Castelo Branco	259	39	2662	1216	1892	198	1909	1705	235	1621	11736
Coimbra	806	254	4993	3219	5264	535	5652	5673	1059	3637	31092
Évora	206	62	1676	1319	1857	168	2001	1683	232	1468	10672
Faro	1161	151	5227	5018	8270	764	9197	5719	672	5064	41243
Guarda	176	15	2960	1004	1379	117	1161	1205	143	1689	9849
Leiria	689	89	4935	3785	5654	557	5989	4806	603	3925	31032
Lisboa	7343	1947	27301	28145	46752	5387	59946	56653	9440	28105	271019
Madeira	378	70	3100	2673	3412	433	3637	4096	416	2578	20793
Portalegre	169	20	1281	805	1219	72	1081	1025	104	950	6726
Porto	3062	706	18235	17649	22806	2212	23397	25252	3824	13770	130913
Santarém	827	87	6307	4702	7272	595	6843	4966	572	5028	37199
Setúbal	2065	265	11981	11125	18359	1615	19908	14218	1590	10941	92067
Viana do Castelo	252	42	4023	2000	2609	198	2207	1854	241	2152	15578
Vila Real	195	30	3413	1722	1948	121	1719	1675	225	1932	12980
Viseu	378	53	5187	3076	3907	262	3210	3009	389	2939	22410

Anexo B : Scripts do R

```
### Dados
dados1<-read.table("Dados.txt",sep="\t",dec=",",header=TRUE)
dados1
rownames(dados1)=dados1[,1]

##Dados sexo
dadosx<-dados1[,2:3]/dados1[,22]
## Dados idade
dadosx<-dados1[,17:21]/dados1[,22]
#Dados trabalho
dadosx<-dados1[,4:6]/dados1[,22]
## Dados habilitação
dadosx<-dados1[,7:16]/dados1[,22]
## Dados Geral
dadosx<-dados1[,2:21]/dados1[,22]

## Medidas metricas
d=dist(dadosx,method="euclidian")

##### método vizinho mais próximo
hc = hclust(d, method="single")
plot(hc,hang=-1,rownames(dados1),col="4",main="vizinho-mais
próximo")
rect.hclust(hc, k=3, border="red")

##### método vizinho mais afastado
hc = hclust(d, method="complete")
plot(hc,hang=-1,col="4",main="vizinho-mais-afastado")
```

```
grupo<-rect.hclust(hc, k=3, border="red")
```

método das média

```
hc = hclust(d, method="average")
plot(hc,hang=-1,rownames(dados1),col="4",main="Metodo average,
Dados Gerais")
rect.hclust(hc, k=2, border="red")
```

método do centróide

```
hc = hclust(d, method="centroid")
plot(hc,hang=-1,col="4",main="metodo de centróide")
rect.hclust(hc, k=4, border="red")
```

método de Ward

```
hc = hclust(d, method="ward.D")
plot(hc,hang=-1,col="4",main=" metodo de ward")
rect.hclust(hc, k=3, border="red")
```

Correlação cofenética

```
)## Nb. o hc depende do método aglomerativo que vai usar
d.coph <- cophenetic(hc
cor(d, d.coph)
```

Metodo divisivos (DIANA)

```
library(cluster)
dv <- diana(d)
print(dv)
dv2 <- cutree(as.hclust(dv), k = 2)
dv2
plot(dv,which =2,main=" Dados trabalho",nmax.lab = 10)
```

Métodos não hierárquicos

metodo k-means

```
library(cluster)

## Determinar o numero k com a regra do cotovelo

k<- 7

wss <- sapply(1:k,

       function(k){kmeans(dadosx,   k,   nstart=50,iter.max   =   7
)$tot.withinss})

wss

plot(1:k, wss,

     type="b", pch =1, frame = FALSE,

     xlab="Number ofClusters s K",

     ylab="Total  within-clusters  sum  of  squares",main="Grafico
Elbow dados sexo" )


###library(amap)

km<-Kmeans(dadosx,2,method="euclidean")

plot(dadosx[c("M","F")],col=km$cluster,main="Kmeans      clsutering
dados sexo",pch=20,cex=2,type="n")

text(dadosx[c("M","F")],col=km$cluster,main="Kmeans      clsutering
dados sexo",pch=20,cex=1,lab=rownames(dados1))

points(km$centers,col="blue",pch=20,cex=1)

abline(v=0.5)

abline(h=0.5)
```

metodo k-medóides (PAM)

```
library(clValid)

## Replicar o PAM para um "k" variando de 2 a 10

pam<-pam(d,2)

plot(pam)


## replicar o número de Clusters  variando de 2 a 8

for(i in 1:8){

  lista.temp[[i]] <- pam(d, i+1)
```

```
 }
## Criando o gráfico de silhueta
par(mfrow=c(2,5))
for(i in 1:8){
 si2<-silhouette(lista.temp[[i]]$cluster,dist(dadosx,"euclidian"))
    plot(si2, do.n.k=FALSE,do.clus.stat=FALSE,cex.names=0.6,
    main=paste("k = ",i+1,sep=""),adj=1)
}
```

####### validação Interna

```
intern<-
 clValid(dadosx,2:6,clMethods=c("hierarchical","diana","kmeans","p
am"),validation="internal",metric="euclidean",method="average")
summary(intern)
op <- par(no.readonly=TRUE)
par(mfrow=c(2,2),mar=c(4,4,3,1))
plot(intern,main="    Validação    interna    Dados    trabalho",
legend=FALSE)
plot(nClusters(intern),measures(intern,"Dunn")[,,1],type="n",axes=
F,xlab="",ylab="")
legend("center",Clusters    Methods(intern),    col=1:9,    lty=1:9,
pch=paste(1:9))
par(op)
```

####### validação de estabilidade

```
stab<-
 clValid(dadosx,2:6,clMethods=c("hierarchical","diana","kmeans","p
am"),validation="stability",metric="euclidean",method="average")
summary(stab)
op <- par(no.readonly=TRUE)
par(mfrow=c(2,3),mar=c(4,4,2,1))
plot(stab,main="Validação etabilidade Dados geral", legend=FALSE)
plot(nClusters(intern),measures(intern,"Dunn")[,,1],type="n",axes=
F,xlab="",ylab="")
legend("center",clusterMethods(intern),col=1:9,lty=1:9,
pch=paste(1:9)
```